CEUR
Workshop
Proceedings
http://ceur-ws.org
ISSN 1613-0073

Vol-1589
urn:nbn:de:0074-1589-9

MultiLingMine 2016
Modeling, Learning and Mining for Cross/Multilinguality

Proceedings of the First Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine 2016) co-located with the 38th European Conference on Information Retrieval (ECIR 2016)

Padova, Italy, March 20, 2016.

Edited by

Dino Ienco *
Mathieu Roche **
Salvatore Romeo ***
Paolo Rosso ****
Andrea Tagarelli *****

* UMR TETIS, IRSTEA, Montpellier, FRANCE
** UMR TETIS, CIRAD, Montpellier, FRANCE
*** QCRI, Doha, Qatar
**** Universitat Politècnica de València, Valencia, Spain
***** Università della Calabria, Rende, Italy

MultiLingMine 2016: Modeling, Learning and Mining for Cross/Multilinguality

Salvatore Romeo[1], Andrea Tagarelli[2], Dino Ienco[3],
Mathieu Roche[4], and Paolo Rosso[5]

[1] Qatar Computing Research Institute, Doha, Qatar
[2] DIMES, University of Calabria, Rende, Italy
[3] IRSTEA, LIRMM, Montpellier, France
[4] CIRAD, LIRMM, Montpellier, France
[5] Universitat Politecnica de Valencia, Valencia, Spain

Abstract. The increasing availability of text information coded in many different languages poses new challenges to modern information retrieval and mining systems in order to discover and exchange knowledge at a larger world-wide scale. The 1st International Workshop on Modeling, Learning and Mining for Cross/Multilinguality (dubbed MultiLingMine 2016) provides a venue to discuss research advances in cross-/multilingual related topics, focusing on new multidisciplinary research questions that have not been deeply investigated so far (e.g., in CLEF and related events relevant to CLIR). This includes theoretical and experimental on-going works about novel representation models, learning algorithms, and knowledge-based methodologies for emerging trends and applications, such as, e.g., cross-view cross-/multilingual information retrieval and document mining, (knowledge-based) translation-independent cross-/multilingual corpora, applications in social network contexts, and more.

1 Motivations

In the last few years the phenomenon of multilingual information overload has received significant attention due to the huge availability of information coded in many different languages. We have in fact witnessed a growing popularity of tools that are designed for collaboratively editing through contributors across the world, which has led to an increased demand for methods capable of effectively and efficiently searching, retrieving, managing and mining different language-written document collections. The multilingual information overload phenomenon introduces new challenges to modern information retrieval systems. By better searching, indexing, and organizing such rich and heterogeneous information, we can discover and exchange knowledge at a larger world-wide scale. However, since research on multilingual information is relatively young, important issues still remain uncovered:

- how to define a translation-independent representation of the documents across many languages;

- whether existing solutions for comparable corpora can be enhanced to generalize to multiple languages without depending on bilingual dictionaries or incurring bias in merging language-specific results;
- how to profitably exploit knowledge bases to enable translation-independent preserving and unveiling of content semantics;
- how to define proper indexing and multidimensional data structures to better capture the multi-topic and/or multi-aspect nature of multi-lingual documents;
- how to detect duplicate or redundant information among different languages or, conversely, novelty in the produced information;
- how to enrich and update multi-lingual knowledge bases from documents;
- how to exploit multi-lingual knowledge bases for question answering;
- how to efficiently extend topic modeling to deal with multi/cross-lingual documents in many languages;
- how to evaluate and visualize retrieval and mining results.

2 Objectives, topics, and outcomes

The aim of the *1st International Workshop on Modeling, Learning and Mining for Cross/Multilinguality* (dubbed *MultiLingMine 2016*),[6] held in conjunction with the 2016 ECIR Conference, is to establish a venue to discuss research advances in cross-/multilingual related topics. MultiLingMine 2016 has been structured as a *full-day* workshop. Its program schedule includes invited talks as well as a panel discussion among the participants. It is mainly geared towards students, researchers and practitioners actively working on topics related to information retrieval, classification, clustering, indexing and modeling of multilingual corpora collections. A major objective of this workshop is to focus on research questions that have not been deeply investigated so far. Special interest is devoted to contributions that aim to consider the following aspects:

- Modeling: methods to develop suitable representations for multilingual corpora, possibly embedding information from different views/aspects, such as, e.g., tensor models and decompositions, word-to-vector models, statistical topic models, representational learning, etc.
- Learning: any unsupervised, supervised, and semi-supervised approach in cross/multilingual contexts.
- The use of knowledge bases to support the modeling, learning, or both stages of multilingual corpora analysis.
- Emerging trends and applications, such as, e.g., cross-view cross-/multilingual IR, multilingual text mining in social networks, etc.

Main research topics of interest in MultiLingMine 2016 include the following:

- Multilingual/cross-lingual information access, web search, and ranking

[6] http://events.dimes.unical.it/multilingmine/

- Multilingual/cross-lingual relevance feedback
- Multilingual/cross-lingual text summarization
- Multilingual/cross-lingual question answering
- Multilingual/cross-lingual information extraction
- Multilingual/cross-lingual document indexing
- Multilingual/cross-lingual topic modeling
- Multi-view/Multimodal representation models for multilingual corpora and cross-lingual applications
- Cross-view multi/cross-lingual information retrieval and document mining
- Multilingual/cross-lingual classification and clustering
- Knowledge-based approaches to model and mine multilingual corpora
- Social network analysis and mining for multilinguality/cross-linguality
- Plagiarism detection for multilinguality/cross-linguality
- Sentiment analysis for multilinguality/cross-linguality
- Deep learning for multilinguality/cross-linguality
- Novel validity criteria for cross-/multilingual retrieval and learning tasks
- Novel paradigms for visualization of patterns mined in multilingual corpora
- Emerging applications for multilingual/cross-lingual domains

The ultimate goal of the MultiLingMine workshop is to increase the visibility of the above research themes, and also to bridge closely related research fields such as information access, searching and ranking, information extraction, feature engineering, text mining and machine learning.

3 Advisory board

The scientific significance of the workshop is assured by a Program Committee which includes 20 research scholars, coming from different countries and widely recognized as experts in cross/multi-lingual information retrieval:
Ahmet Aker, Univ. Sheffield, United Kingdom
Rafael Banchs, I2R Singapore
Martin Braschler, Zurich Univ. of Applied Sciences, Switzerland
Philipp Cimiano, Bielefeld University, Germany
Paul Clough, Univ. Sheffield, United Kingdom
Andrea Esuli, ISTI-CNR, Italy
Wei Gao, QCRI, Qatar
Cyril Goutte, National Research Council, Canada
Parth Gupta, Universitat Politcnica de Valncia, Spain
Dunja Mladenic, Jozef Stefan International Postgraduate school, Slovenia
Alejandro Moreo, ISTI-CNR, Italy
Alessandro Moschitti, Univ. Trento, Italy; QCRI, Qatar
Matteo Negri, FBK - Fondazione Bruno Kessler, Italy
Simone Paolo Ponzetto, Univ. Mannheim, Germany
Achim Rettinger, Institute AIFB, Germany
Philipp Sorg, Institute AIFB, Germany
Ralf Steinberger, JRC in Ispra, Italy
Marco Turchi, FBK - Fondazione Bruno Kessler, Italy

Vasudeva Varma, IIIT Hyderabad, India
Ivan Vulic, KU Leuven, Belgium

4 Related events

A COLING'08 workshop [1] was one of the earliest events that emphasized the importance of analyzing multilingual document collections for information extraction and summarization purposes. The topic also attracted attention from the semantic web community: in 2014, [2] solicited works to discuss principles on how to publish, link and access mono and multilingual knowledge data collections; in 2015, another workshop [3] took place on similar topics in order to allow researchers continue to address multilingual knowledge management problems. A tutorial on Multilingual Topic Models was presented at WSDM 2014 [4] focusing on how statistically model document collections written in different languages. In 2015, a WWW workshop aimed at advancing the state-of-the-art in Multilingual Web Access [5]: the contributing papers covered different aspects of multilingual information analysis, leveraging attention on the lack of current information retrieval techniques and the necessity of new techniques especially tailored to manage, search, analysis and mine multilingual textual information.

The main event related to our workshop is the CLEF initiative [6] which has long provided a premier forum for the development of new information access and evaluation strategies in multilingual contexts. However, differently from MultiLingMine, it does not have emphasized research contributions on tasks such as searching, indexing, mining and modeling of multilingual corpora.

Our intention is to continue the lead of previous events about multilingual related topics, however from a broader perspective which is relevant to various information retrieval and document mining fields. We aim at soliciting contributions from scholars and practitioners in information retrieval that are interested in Multi/Cross-lingual document management, search, mining, and evaluation tasks. Moreover, differently from previous workshops, we would emphasize some specific trends, such as cross-view cross/multilingual IR, as well as the growing tightly interaction between knowledge-based and statistical/algorithmic approaches in order to deal with multilingual information overload.

References

1. Bandyopadhyay, S., Poibeau, T., Saggion, H., Yangarber, R. (2008). Procs. of the Workshop on Multi-source Multilingual Information Extraction and Summarization (MMIES). ACL.
2. Chiarcos, C., McCrae J. P., Montiel E., Simov, K., Branco, A., Calzolari, N., Osenova, P., Slavcheva, M., Vertan, C. (2014). Procs. of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and NLP (LDL).
3. McCrae, J. P., Vulcu G. (2015). CEUR Procs. of the 4th Workshop on the Multilingual Semantic Web (MSW4), Vol. 1532.

4. Moens, M.-F., Vulié, I. (2014). Multilingual Probabilistic Topic Modeling and Its Applications in Web Mining and Search. In Procs. of the 7th ACM WSDM Conf.
5. Steichen, B., Ferro, N., Lewis, D., Chi, E. E. (2015). Procs. of the Int. Workshop on Multilingual Web Access (MWA).
6. The CLEF Initiative. http://www.clef-initiative.eu/.

Identification of Disease Symptoms in Multilingual Sentences: an Ontology-Driven Approach*

Angelo Ferrando[1], Silvio Beux[1], Viviana Mascardi[1], and Paolo Rosso[2]

[1]DIBRIS, Università degli Studi di Genova, Italy
angelo.ferrando@dibris.unige.it, silviobeux@gmail.com,
viviana.mascardi@unige.it
[2]PRHLT, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract. In this paper we present a Multilingual Ontology-Driven framework for Text Classification (MOoD-TC). This framework is highly modular and can be customized to create applications based on Multilingual Natural Language Processing for classifying domain-dependent contents. In order to show the potential of MOoD-TC, we present a case study in the e-Health domain.

Key words: Multilingual Natural Language Processing, Ontology-Driven Text Classification, BabelNet, Symptom Disease Identification

1 Introduction

The large amount of digital data made available in the last years from a wide variety of sources raises the need for automatic methods to extract meaningful information from them. The extracted information is precious for many purposes, and especially for commercial ones. Jackson and Moulinier [12] observe that *"there is no question concerning the commercial value of being able to classify documents automatically by content. There are myriad potential applications of such a capability for corporate Intranets, government departments, and Internet publishers"*.

The problem of classifying multilingual pieces of text was addressed since the end of the last millennium [17] but it is still a significant problem because each language has its own peculiar features, making the automatic management of multilingualism an open issue.

The use of ontologies to classify multilingual texts [5] is a good alternative to standard machine learning approaches in all those situations where a training set of documents is not available or it is too small to properly train the classifier. Ontology-driven text classification does not depend on the existence of a training set, as it relies solely on the entities, their relationships, and the taxonomy of categories represented in an ontology, that becomes the driver of the

* The first author of this paper is a PhD student in Computer Science at the University of Genova, Italy. The work of the last author was in the framework of the SomEMBED MINECO TIN2015-71147-C2-1-P research project.

classification. Another advantage of ontology-driven classification is that ontology concepts are organized into hierarchies and this makes possible to identify the category (or the categories) that best classify the document's content, by traversing the hierarchical structure.

In this paper we present MOoD-TC (*M*ultilingual *O*ntology *D*riven *T*ext *C*lassifier [3, 13]), a highly modular system which has been conceived, designed and implemented to be customized by the system developer for obtaining different domain-dependent behaviors, always centered around the multilingual text classification process. The original contribution of this paper is the exploitation of the core "multilingual word identification" functionalities of MOoD-TC for a challenging scenario in the e-Health domain, where classification is a by-product of disease symptoms identification in multilingual pieces of text, driven by a standard symptoms ontology. A customization of MOoD-TC with an ad-hoc module equipped with pre- and post-processing facilities suitable for the scenarios that motivate our work, is also described.

The paper is organized as follows: Section 2 introduces three motivating scenarios where an ontology-driven multilingual text classification may prove useful, Section 3 analyzes the state of the art, Section 4 describes MOoD-TC, Section 5 provides examples and experimental results, and Section 6 concludes.

2 Motivating scenarios

Alice is enjoying her holidays in Stockholm. Suddenly, she feels a painful spasm to her stomach and in a few minutes a strong feeling of nausea appears. Spasms go on for half an hour, and she starts to feel worried. She does not think it is the case to go to the hospital, but she would at least ask for advice over the phone. However, she cannot speak Swedish and, in the stressful situation she is experiencing, she cannot recall how to express her health problems in English. She could speak in her native language Italian, but it is not so likely that the doctor can speak Italian as well.

Bob is making a walk in his town. He notices a young man bending over his knees, with a scared expression on his face. He runs to help him, and he understands that the problem is with his chest. The young man speaks French only and Bob cannot understand him: he calls the first aid emergency number and explains what he is seeing and what he supposes to be taking place. If he could understand what the young man says, he would be definitely more helpful.

Carol is a volunteer in Honduras. She is neither a physician nor a nurse. She has a very basic knowledge of first aid procedures and a first aid kit with medicines that she knows how to administer, given a clear diagnosis. A woman runs towards her asking for her assistance. The woman's small boy has a problem with his head and he has a high fever but, without understanding the other symptoms that the woman is trying to explain in Spanish, Carol cannot recognize and classify the problem. In the remote place where she is, she cannot contact the doctor. Carol should need to understand the other symptoms besides fever and headache, in order to select the correct medicine.

The three scenarios above are all characterized by the impossibility for the doctor to visit the patient on-the-fly and the need for the patient to be understood despite language barriers, in order to get advice for minor problems or to speed up the assistance procedure for major ones. The patient's need could be suitably addressed by identifying and translating symptoms from her language to the assistant's or the doctor's one. If automatic tools for facing this issue were available, for example as an app installed on the mobile phone, the three situations could evolve in the following way:

- **Scenario 1**: through the use of an app, the person needing care communicates with the "health emergency" software application in her own language. The application performs a speech-to-text translation, **identifies the symptoms in the text based on a standard ontological representation of symptoms**, and sends the list of symptoms expressed in the doctor's language to a center where they are managed either by intelligent software agents or by human experts.
- **Scenario 2**: the "health emergency" software application is not directly used by the person needing care, but by the one who assists her. Like before, the assisted person can "tell" her problems to the application which performs a speech-to-text translation and **identifies the symptoms represented in a domain ontology which appear in the text**. The symptoms, translated into the language of the person who his giving the first assistance, may be read on the screen. That person can call the national first aid number, telling what is happening, what she sees, and the symptoms which have been understood, classified, and translated by the app.
- **Scenario 3**: also in this case, besides a speech-to-text translation, **the symptoms expressed in the language of the patient are identified w.r.t. a symptoms ontology** and translated into the target language. The way this information is used can require a further automatic processing stage, if the doctor cannot be involved in the loop and the person providing aid needs an automatic support for making a diagnosis and identifying the right therapy to administer.

In all the three situations above, a standard machine translation application and a symptoms classifier based on machine learning might not be powerful enough: the pre- and post-processing stages require to have a machine-readable explicit representation of symptoms, in some vocabulary agreed upon by all the application components and by the humans involved in the loop, in order to share them among the application components (both at the client and at the server side) and to reason about them if needed. A multilingual ontology-driven text classification approach seems the right way to face these challenging scenarios.

3 State of the art

According to [8], in 1996 more than 80% of Internet users were native English speakers. This percentage has dropped to 55% in 2000 and to 27.3% in 2010. However, about 80% of the digital resources available today on the Web (including deep Web and digital libraries) are in English [10]. This calls for the urgent need of establishing multilingual information systems and Cross-Language Information Retrieval (CLIR) facilities. How to manipulate the large volume of multilingual data has now become a major research question.

In this paper we are interested in Natural Language Processing (NLP) techniques for solving multilingual term identification and text classification problems in the e-Health domain where extracting information from clinical notes has been the focus of a growing body of research in the past years [14]. Common characteristics of narrative text used by physicians in electronic health records make the automatic extraction of meaningful information hard. NLP techniques are needed to convert data from unstructured text to a structured form readily processable by computers [15]. This structured representation can be used to extract meaning and enable Clinical Decision Support systems that assist healthcare professionals and improve health outcomes [6].

Signs and symptoms have seldom been studied for themselves in the field of biomedical information extraction. Indeed, they are often included in more general categories such as "clinical concepts" [22], "medical problems" [21] or "phenotypic information" [19]. Moreover, most of the available studies are based on clinical reports or narrative corpora. In [11, 18], indeed, the aim consists in symptom extraction from clinical records and in [20] the authors identify the risk factors for heart disease based on the automated analysis of narrative clinical records of diabetic patients.

Another recent project in e-Health NLP context is the IBM Watson for Oncology[1]. It has an advanced ability to analyze the meaning and context of structured and unstructured data in clinical notes and reports, easily assimilating key patient information written in plain English that may be critical to select a treatment pathway. These works are different from ours because they do not address multilingual aspects and, furthermore, because they have to manage the differences between the "signs", which are identified by clinicians, and the "symptoms", which can be described directly by the sick person.

In our work we do not have to manage clinical records but directly the information provided by the person who feels sick. This difference is crucial in works using an ontology-driven approach, because clinical reports contain many more technical words[2] compared to a text written (or a sentence told) by a normal person describing how she feels. This allows us to use simpler ontologies. Especially from the multilingual viewpoint, having an ontology containing simple concepts, omitting useless technicalities, allows achieving better results with less effort, considering that a technical word could be less supported by the tools we use during our text classification pipeline.

The assumption upon which MOoD-TC relies, is the availability of ontologies in the domain of interest. Even if the application developer might design and implement her own domain ontology from scratch, integrating well-founded and widely used ontologies into MOoD-TC would be the most modular, reusable and scientifically acceptable approach. Luckily, many domain ontologies exist, in particular in the biomedical domain. Panacea [7], the Ontology for General Medical Science[3], and the Gene Ontology[4] are just a few recent examples, besides the "symptoms ontology" used for our experiments and discussed in Section 5.

[1] http://www.ibm.com/smarterplanet/us/en/ibmwatson/watson-oncology.html

[2] A clinical report is written by a doctor.

[3] https://bioportal.bioontology.org/ontologies/OGMS

[4] http://geneontology.org/

4 MOoD-TC

MOoD-TC has been developed as part of Silvio Beux' Masters Thesis [3], starting from [13]. Its aim is to classify multilingual textual documents according to classes described in a domain ontology. MOoD-TC consists of the Text Classifier (TC) and the Application Domain Module (ADM). It provides a set of core modules offering functionalities which are common to any text classification problem (text pre-processing, tagging, classification) plus a customizable structure for those modules which can be implemented by the developer in order to offer application-specific functionalities. It returns a classification of the text w.r.t. the ontology taken as input. The classification performed by TC which is implemented in Java and exploits the Language Detector Library[5], BabelNet [16], and TreeTagger[6].

The Language Detector Library detects, with a precision greater than 99%, 53 languages making use of Naive Bayesian filters. It is devoted to recognize the language L_o of the ontology o and the language L_d of the textual document d. The TreeTagger tool performs the tagging of d in order to obtain, for each word $w \in d$ different from a stop word, its lemma (the canonical form of the word) and its part of speech (POS). This information is used by BabelNet to perform the translation of w into the ontology language. Finally, the translated word w' is searched inside the ontology and contributes to the classification of d in the category modeled by the ontology concept c having the same semantics as w'. The *ClassifierObject* is the object that stores a correctly classified word (and additional information) of the document d with respect to o. TC returns a list of such objects. ADM specializes the text classifier task by implementing functionalities for pre- and post- processing a multilingual textual document. If an ADM is used, the entire system specializes its behaviour in the domain represented by that particular ADM (e.g., from text classifier to disease recognizer). In our system TC can work alone, but an ADM is meant to work in close connection with the core system. The core modules are implemented to work for the European languages (which share some common features like, for example, the relationship between noun and adjective), but they could be extended to cope with the peculiar features of other languages; in fact, thanks to the modularity of the system, it is possible to integrate different algorithms created specifically to handle that peculiarities, without modifying the entire system. The ADM processes the TC input and output in order to obtain a new domain oriented tool. An ADM is composed by two sub-components: pre-processing and post-processing. The pre-processing component takes as input a digital object (for

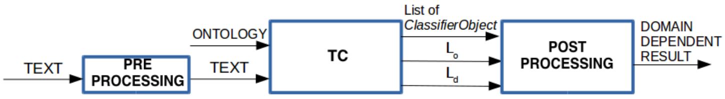


Fig. 1. Integration pipeline of TC and ADM.

[5] https://code.google.com/p/language-detection/

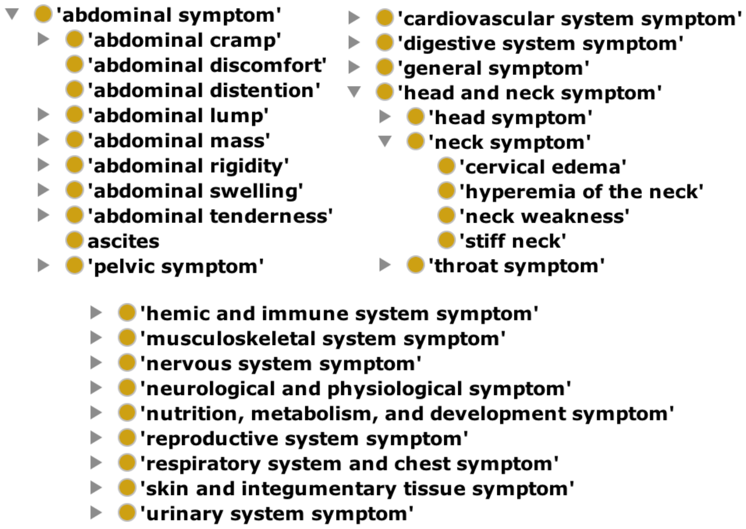
[6] http://code.google.com/p/tt4j/

example a spoken sentence, in the scenarios discussed in Section 2) and returns a new processed text, while the post-processing component takes as input the TC output and returns a domain dependent result. Figure 1 shows the entire pipeline of the integration process between the TC and the ADM.

5 Exploiting MOoD-TC for Symptom Identification

As illustrated in Section 2, the scenarios we aim to address require that disease symptoms appearing in a text are correctly identified w.r.t. a domain ontology. The pre-processing stage consists of moving from a spoken sentence to a text and the post-processing in translating the identified symptoms into a target language and, depending on the scenario, moving back from text to speech and/or reasoning over them. In the sequel we discuss the experiments related with our main task, namely that of symptoms identification.

The domain ontology used for describing symptoms is a standard ontology named the *symptoms ontology*[7], partially shown in Figure 2. It is an ontology of disease symptoms with symptoms encompassing perceived changes in function, sensations or appearance reported by a patient and indicative of a disease. We stress that our experiments in exploiting MOoD-TC for symptom identification did not require to build any new ontology. Rather, consistently with the good principle of reusing existing software whenever available and, in particular, reusing existing ontologies, we just passed the symptoms ontology as input to the TC, obtaining the results discussed in the next section.



- 'abdominal symptom'
  - 'abdominal cramp'
  - 'abdominal discomfort'
  - 'abdominal distention'
  - 'abdominal lump'
  - 'abdominal mass'
  - 'abdominal rigidity'
  - 'abdominal swelling'
  - 'abdominal tenderness'
  - ascites
  - 'pelvic symptom'
- 'cardiovascular system symptom'
- 'digestive system symptom'
- 'general symptom'
- 'head and neck symptom'
  - 'head symptom'
  - 'neck symptom'
    - 'cervical edema'
    - 'hyperemia of the neck'
    - 'neck weakness'
    - 'stiff neck'
  - 'throat symptom'
- 'hemic and immune system symptom'
- 'musculoskeletal system symptom'
- 'nervous system symptom'
- 'neurological and physiological symptom'
- 'nutrition, metabolism, and development symptom'
- 'reproductive system symptom'
- 'respiratory system and chest symptom'
- 'skin and integumentary tissue symptom'
- 'urinary system symptom'

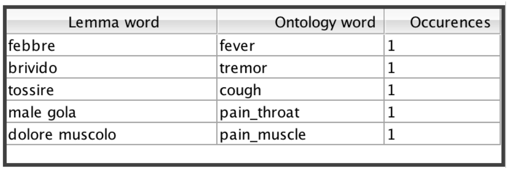
Fig. 2. Symptoms ontology (the three branches are children of "Symptom").

In the sequel we discuss our initial experiments with phrases in five different languages (English, French, German, Italian, Spanish), where symptoms are

[7] http://purl.obolibrary.org/obo/symp.owl

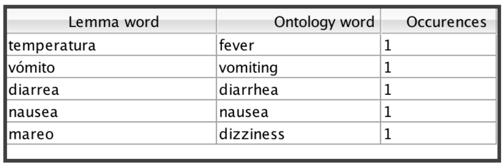
identified by the TC module. The classification of two sample sentences is shown below, where the TC GUI screenshot associated with each sentence shows the ontology concepts which appear in the text along with the number of their occurrences in the text.

Phrase 1 (Italian language): *"Credo di avere la febbre, continuo a sudare e ho i brividi. Non la smetto di tossire e fatico a mangiare a causa del male alla gola, come un forte bruciore. Mi sento stanchissimo e ho dolore a tutti i muscoli."*



Lemma word	Ontology word	Occurences
febbre	fever	1
brivido	tremor	1
tossire	cough	1
male gola	pain_throat	1
dolore muscolo	pain_muscle	1

Phrase 3 (Spanish language): *"Me siento fatal. Tengo temperatura, vòmito y diarrea. Hace dos dìas que no consigo comer nada. Tengo nausea y mareos."*



Lemma word	Ontology word	Occurences
temperatura	fever	1
vómito	vomiting	1
diarrea	diarrhea	1
nausea	nausea	1
mareo	dizziness	1

The experiments have been carried out on 32 sentences for each of the 5 languages, for a total of 160 sentences. Each sentence describes symptoms related to one of the following sixteen disease: tinnitus, food allergy, cervical, dehydration, hyperthyroidism, flu, appendicitis, food poisoning, labyrinthitis, narcolessia, pneumonia, diabetes type 1, hyperglycemia, hypoglycemia, bronchitis, jet lag (two sentences for each disease). To cover the widest range of cases we considered the diseases with the most varied symptoms. The description of symptoms associated with each disease has been retrieved from [9] and each sentence contains 2 up to 9 symptom words. The sentences were manually created by the authors.

Since the final purpose of this work is to provide an automatic diagnostic system with as many symptoms as possible, in order to devise the correct diagnosis, we were mainly interested in symptoms which appear in the text but which are not identified by our classifier (false negatives). We also looked for false positives, but their number is so low to be irrelevant for our experiments. Also, false positives are due to an under classification, rather than an actually wrong classification: if the text contains the "abdominal cramp" symptom, for example, and it is classified with the more general "abdominal symptom" concept, we consider this result a false positive as a more specific concept could have been returned. Figure 3 shows the average number of symptoms that should have been identified w.r.t the correctly identified symptoms in the five considered languages. Figure 4 shows the number of false negatives (y axis) for disease (x axis). Figure 3 demonstrates that the results greatly vary with the disease. For example, symptoms related to tinnitus are hardly classified, but this can be easily explained by observing the ontology we used, where problems related to ears are not modeled at all. By carefully analyzing the obtained results, we also realized that sometimes the performances of the classifier are worsened by the presence of a symptom in the text which has a different grammatical role than the symptom in the ontology (usually a noun), making their matching impossible although the word root and the meaning are the same. For example, the ontology contains the noun "irritability", but if the text contains the adjective "irritable" (in any

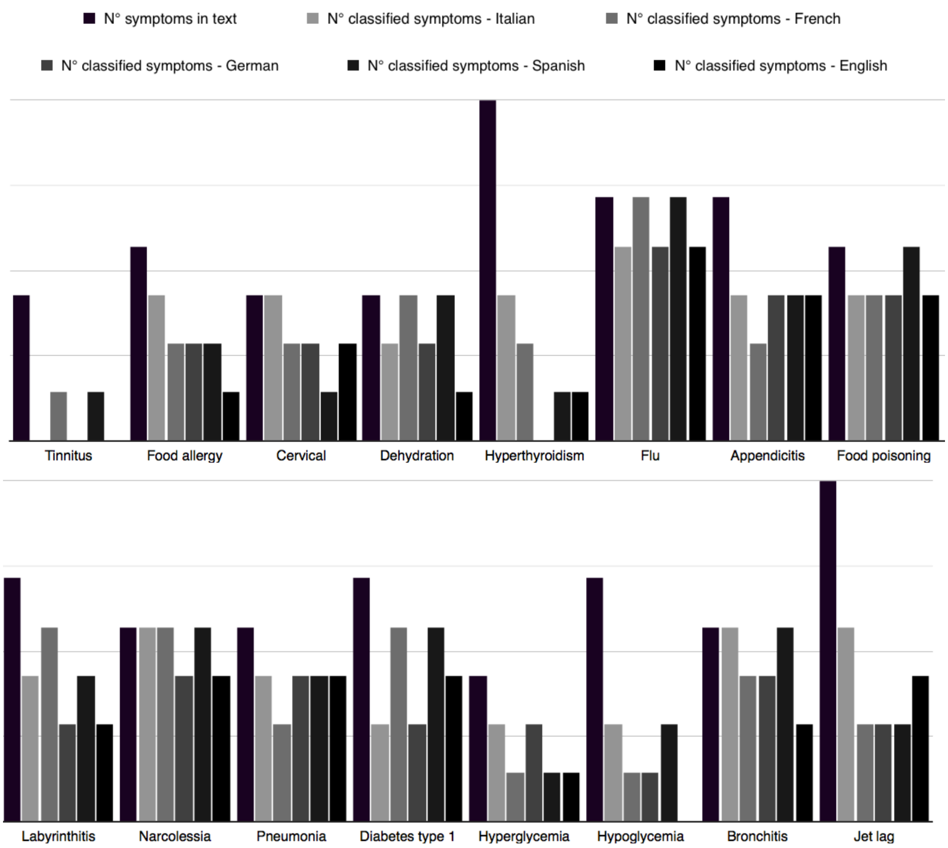


Fig. 3. For each disease, the leftmost column (in black) measures the average number of symptoms that should have been identified; the next five columns show the average number of correctly identified symptoms in Italian, French, German, Spanish and English sentences respectively.

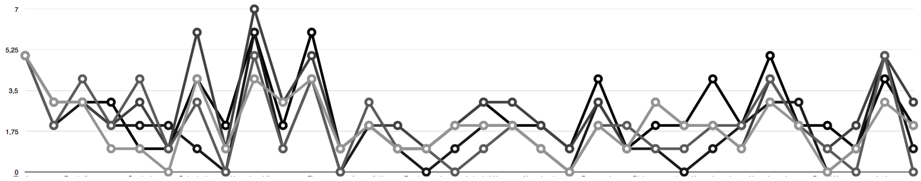


Fig. 4. Trend of errors for disease in the five languages (False Negatives). On the x axis the diseases are reported (labels are omitted) and on the y axis the number of false negatives for disease is reported: each line in the graphic is associated with one language.

language), the identification fails. This problem is due to the way the root of a word is computed, and to the way words are managed in BabelNet.

What emerges from Figure 4 is that false negatives have a very similar behavior despite the language of the sentence. This is again due to the two reasons discussed above. Despite these problems, which have a clearly understood motivation and which can be addressed by extending the ontology and by refining the management of word root extraction, MOoD-TC has demonstrated to be a flexible and ready-to-use approach for multilingual symptoms identification driven by a standard ontology we retrieved on the web.

6 Conclusions and Future Work

In this paper we presented the MOoD-TC architecture showing its possible use in the symptoms identification problem. The speech-to-text pre-processing stage can be faced using existing tools, and the post-processing stage with a translation of the identified symptoms into the doctor's language can be addressed using BabelNet, in the same way we exploit BabelNet for bridging the text, whatever its language, and the ontology. The more challenging post-processing stage of supporting the user in providing a diagnosis given a set of identified symptoms could be addressed by means of sophisticated expert system such as the old and well known MYCIN [4] and more recent projects (http://www.easydiagnosis.com/, https://www.diagnose-me.com/, [2]), some of which are ontology-driven [1].

Our framework does not face many well known open problems in multilingual text classification and information extraction such as negation [23] and named entities, but rather it provides a flexible and modular approach ready for integrating, with limited effort, the results and algorithms addressing the above problems coming from the research community.

In the short time, our work will be devoted to overcome the problems that limit the performances of MOoD-TC in the considered scenario: we will make the word identification more flexible and we will extend the symptoms ontology with those symptoms which have not been modeled so far.

In the future, it would be interesting to run an experimental comparison between our approach and a machine learning one. In case of a limited number of labeled examples, in fact, it would be feasible to apply semi-supervised learning methods. Depending on the comparison results, we will also consider to combine both approaches, using a domain ontology to improve the results of a traditional machine learning approach.

References

1. B. Al-Hamadani. CardioOWL: An ontology-driven expert system for diagnosing coronary artery diseases. In *2014 IEEE Conference on Open Systems (ICOS)*, pages 128–132, 2014.
2. R. P. Ambilwade, R. R. Manza, and B. P. Gaikwad. Medical expert systems for diabetes diagnosis: A survey. *Int. J. of ARCSSE*, 4(11), 2014.
3. S. Beux. MOoD-TC: A general purpose multilingual ontology driven text classifier. Master's Degree Thesis in Computer Science, University of Genova, Italy, 2015.
4. B. G. Buchanan and E. H. Shortliffe. *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, 1984.

5. G. de Melo and S. Siersdorfer. Multilingual text classification using ontologies. In *ECIR Conference, Proceedings*, volume 4425 of *LNCS*, pages 541–548. Springer, 2007.
6. D. Demner-Fushman, W. Chapman, and C. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009.
7. C. Doulaverakis, G. Nikolaidis, A. Kleontas, and I. Kompatsiaris. Panacea, a semantic-enabled drug recommendations discovery framework. *J. Biomedical Semantics*, 5:13, 2014.
8. Global Reach. Global internet statistics (by language). Technical report, Global Reach, June 2005.
9. H. W. Griffith. *Complete guide to symptoms, illness & surgery for people over 50.* Body Press/Perigee New York, NY, 1992.
10. B. Guo-Wei and C. Hsin-Hsi. Cross-language information access to multilingual collections on the Intenet. *Journal of the American Society for Information Science*, 51, 2000.
11. H. Harkema, I. Roberts, R. Gaizauskas, and M. Hepple. Information extraction from clinical records. In S. Cox, editor, *Proceedings of the 4th UK e-Science All Hands Meeting*, Nottingham, UK, 2005.
12. P. Jackson and I. Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction & Categorization.* John Benjamins, 2002.
13. M. Leotta, S. Beux, V. Mascardi, and D. Briola. My MOoD, a multimedia and multilingual ontology driven MAS: design and first experiments in the sentiment analysis domain. In *ESSEM Workshop, Proceedings*, pages 51–66, 2015.
14. S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, pages 128–144, 2008.
15. P. Nadkarni, L. Ohno-Machado, and W. Chapman. Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.
16. R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.
17. D. W. Oard and B. J. Dorr. A survey of multilingual text retrieval. Technical report, College Park, MD, USA, 1996.
18. A. Rosier, A. Burgun, and P. Mabo. Using regular expressions to extract information on pacemaker implantation procedures from clinical reports. In *Proceedings of the AMIA Annual Symposium*, Washington DC, USA, 2008.
19. B. R. South, S. Shen, M. Jones, J. H. Garvin, M. H. Samore, W. W. Chapman, and A. V. Gundlapalli. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *BMC Bioinformatics*, 10(S-9):12, 2009.
20. A. Stubbs, C. Kotfila, H. Xu, and Özlem Uzuner. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of Biomedical Informatics*, 58, Supplement:S67 – S77, 2015.
21. O. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 18(5):552–556, 2011.
22. K. B. Wagholikar, M. Torii, S. Jonnalagadda, H. Liu, et al. Pooling annotated corpora for clinical concept extraction. *J. Biomedical Semantics*, 4:3, 2013.
23. M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 60–68, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Deep Level Lexical Features for Cross-lingual Authorship Attribution.

Marisa Llorens-Salvador. Sarah Jane Delany.

Dublin Institute of Technology, Dublin, Ireland

Abstract. Crosslingual document classification aims to classify documents written in different languages that share a common genre, topic or author. Knowledge-based methods and others based on machine translation deliver state-of-the-art classification accuracy, however because of their reliance on external resources, poorly resourced languages present a challenge for these type of methods. In this paper, we propose a novel set of language independent features that capture language use from a document at a deep level, using features that are intrinsic to the document. These features are based on vocabulary richness measurements and are text length independent and self-contained, meaning that no external resources such as lexicons or machine translation software are needed. Preliminary evaluation results show promising results for the task of crosslingual authorship attribution, outperforming similar methods.

Keywords: Crosslingual document classification, crosslingual authorship attribution, deep level lexical features,vocabulary richness features.

1 Introduction

Despite the prevalence of the English language in many fields, international organizations manage large numbers of documents in different languages, from local legislation to internal documents produced in different company locations. At the same time, workers' mobility has created a multilingual work force that create and store documents in different languages depending on the context. For example, the same author can write academic papers in English, write a technical book in French and a novel in Catalan. The classification of these multilingual documents has applications in the areas of information retrieval, forensic linguistics and humanities scholarship.

The analysis of document style and language use has long been used as a tool for author attribution. Traditionally, research in the area focused on monolingual corpora [12] or employed external resources such as machine translation, multilingual lexicons or parallel corpora [3, 14, 15].

In this paper, we present a set of language independent lexical features and study their performance when used to solve the problem of crosslingual author attribution. The task of crosslingual author attribution (CLAA) refers to the identification of the author of a document written in language x_i from a pool of known authors whose known documents are written in languages $x_1, x_2, .., x_n$. The aim of the method is to identify the author of an unseen document without prior knowledge about its language, i.e. without using any language specific features, tuning for a particular language or the use of machine translation/lexicon aid in a completely language independent implementation.

The proposed method builds on traditional vocabulary richness measures (VR), such as type-token ratio or hapaxes frequency. Traditional vocabulary richness features are text-length dependent and provide a small number of features (type-token ratio being the best example with only one value representing each text). In order to overcome these limitations, our proposed method for feature extraction calculates features on fixed length samples of text extracted from the document. Mean and dispersion values for vocabulary richness are calculated obtaining 8 deep level lexical features. The performance of different sample sizes i is studied individually and as combinations of sizes, providing information about text consistency through the document and characteristic vocabulary use.

2 Related Work

Monolingual author attribution has in the last few years achieved a high level of accuracy using lexical features such as frequencies of the most common words and Burrow's Delta to calculate distances between documents [1, 4, 7, 11, 13]. Other lexical features used in monolingual author attribution include frequencies of stop words [2] and word n-grams. In these models, a feature vector with all features (n-grams or stop words) contained in the document and their frequencies characterizes each document. The problem when extending these methods to multilingual corpora is that the dimensions of the feature vectors in different languages are in general orthogonal, giving zero as the similarity measure between documents. Character n-grams have been applied to different languages and have obtained high levels of accuracy at the expense of high dimensionality with feature set sizes in the thousands [7]. At a syntactic level, features such as part-of-speech and frequency of verbs and pronouns have achieved high level of accuracy as well [6]. However, all the above features are either language dependent or involve high dimensional feature sets.

Traditional vocabulary richness like the type-token ratio are language independent, however, they depend on text length and for this reason have been replaced in recent times by more complex features. These features include the Moving Window Type-Token Ratio and the Moving Window Type-Token Ratio Distribution [5, 8]. Despite their language independence nature, traditional measurements of vocabulary richness have not delivered highly accurate results in the past [13]. Consequently, they have been replaced by the use of lexical features in combination with machine translation software or lexicons/dictionaries to bring all documents into the same language space with *wikipedia* and the *eurovoc corpus* the most commonly used resources [9,10,14].

3 Methodology

Based on vocabulary richness and frequency spectrum values, the proposed features and method for feature extraction define a way of quantifying the style of a text by analysing the use of vocabulary in samples of different sizes taken from the text. These samples are based on the idea of a moving window type-token ratio using fixed size samples and hence avoiding the shortcomings of the type-token ratio. These features extend the moving window type-token ratio as more granular measurements of word frequencies are extracted.

Three sampling methods are included in the framework: (i) Fragment sampling (FS), (ii) Bags-of-words sampling (BS) and (iii) the combination of both Bag-Fragment sampling (BFS).

Fragment sampling (FS) is defined as the process of randomly obtaining n samples of i consecutive words, starting from a word chosen at random and each sample is referred to as a fragment. Given the random nature of the sampling process these fragments can overlap and are not following any sequence in terms of overall location in the text. Bags-of-words sampling (BS) involves the use of i words sampled randomly from any part of the document and follows the well known concept of treating a text as a bag-of-words where the location of words is ignored .

The proposed set of language independent lexical features is extracted following a 4 step process:

STEP 1: A number n of document samples of size i is extracted.
STEP 2: Values for frequency features are calculated per sample.
STEP 3: Values for mean and dispersion features calculated across the n samples.
STEP 4: Back to step 1 for a new sample size i.

The general parameters of the method are: type of sample (Fragment, Bags-of-words or both), sample sizes i_1,i_2,...,i_M and number of samples n per sample size. Figure 1 depicts a diagram for the extraction process for BFS. FS and BS are represented by the left and right hand-side of the diagram respectively.

The proposed set of frequency features are based on the analysis of the frequency spectrum, i.e. how many times each feature appears. A typical example of this type of features is the number of hapaxes or words that appear only once in the text. Instead of using the entire frequency spectrum and in order to reduce the number of features and capture information in a compact way, a novel

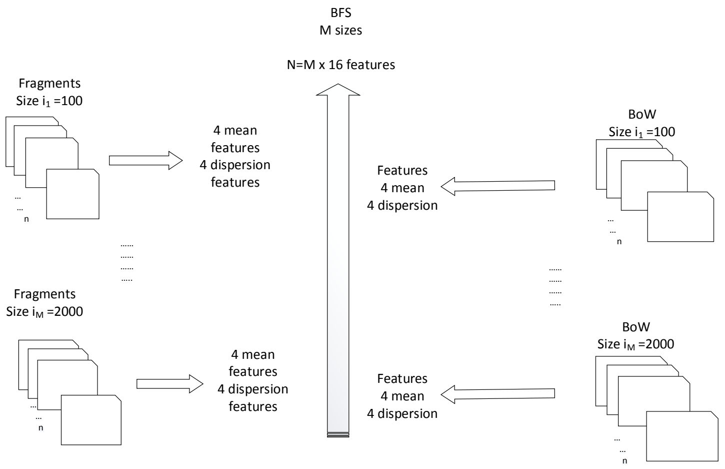


Fig. 1. BFS process summary diagram.

method of frequency spectrum representation is presented.

The frequency spectrum for different texts shows regular behaviour for the initial low frequencies, however, after frequency 10 the number of words for each frequency becomes less stable as can be seen in Figure 2, which shows the frequency spectrum for Charles Dickens' Oliver Twist in its original language. For this reason, frequency values over 10 are not used for the purpose of feature extraction. Notwithstanding these considerations, the words included in that frequency range (over 10) are not entirely neglected as they feature as part of the overall vocabulary and hence contribute to the classification process.

The frequency spectrum for values of frequency between 1 and 10 is regular (quasi linear) and hence suitable for a small number of points to represent its behaviour. In order to reduce the dimensions of the feature set and given the quasi linear behaviour of the data, a further simplification is performed and groupings of 1, 2-4, and 5-10 are used. Each frequency range is represented by a feature, obtaining 3 features to represent the frequency spectrum between 1 and 10 and a separate fourth feature that represents the vocabulary or different unique words present in the text. Figure 2a shows the 3 features representation of data for Charles Dickens' Oliver Twist in its original language English plotted on top of the overall frequency spectrum.

The feature representation of the frequency spectrum for values of frequency between 1 and 10 holds for fragments and bags-of-words samples as shown on Figure 2b. The sampling process allows for dispersion features to be calculated providing a measurement of the homogeneity of the text.

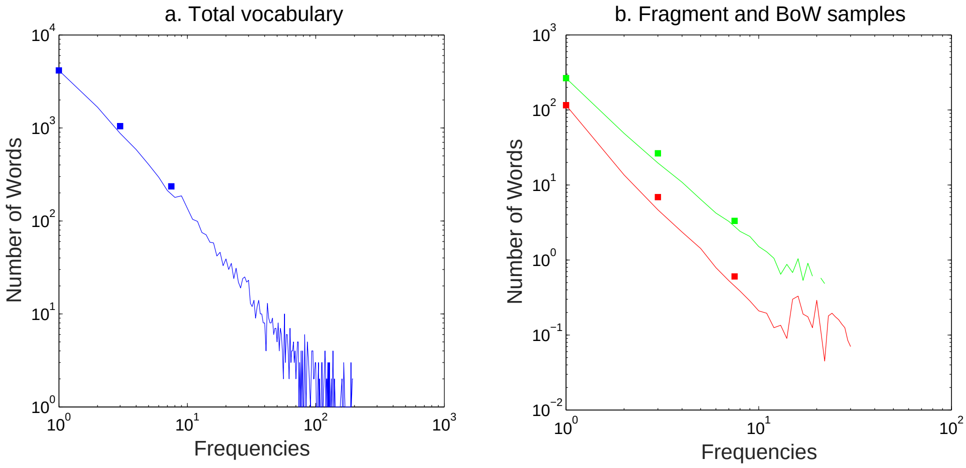


Fig. 2. Oliver Twist (Charles Dickens) a. Frequency spectrum with 3 features b. Fragment and bags-of-words sample (i=200) with 3 features.

Table 1 shows the proposed mean and dispersion features for the frequency groupings and vocabulary.

1	Size of vocabulary per sample.
2	Number of local hapaxes h_i.
3	Number of words with frequency 2, 3 and 4.
4	Number of words with frequency 5 to 10.
5.	Coefficient of variation for vocabulary.
6.	Coefficient of variation for local hapaxes.
7.	Coefficient of variation for words with frequency 2, 3 and 4.
8.	Coefficient of variation for words with frequency 5 to 10.

Table 1. Deep level features.

The sampling process is repeated for a number, M, of sample sizes, i, and the 8 features calculated for each size. This provides a variable number of final features depending on the number of sizes selected. The size of the resulting set of features depends on M, the number of different sizes sampled. The total number of features N is $N = 8 \times M$ for FS and BS and $N = 16 \times M$ for BFS.

3.1 Datasets

In order to adjust the parameters of the proposed feature extraction method, a multilingual corpus of literary works was compiled. Due to the cross-lingual nature of the experiments, documents in different languages created by the same author are required. Literary translation is believed to keep the markers from the original author and the influence of the translator is weak [16], therefore the corpus used in the experiments is formed by original works by 8 authors and translated novels from the same 8 authors. It includes two datasets: Dataset 1, a balanced dataset of original documents and Dataset 2 a unbalanced extended version including translations. Dataset 1 contains 120 literary texts from 8 different authors (15 documents per author) in 4 different languages (English, Spanish, German and French) as can be seen in Table 2. Dataset 2 includes all documents from Dataset 1 plus 85 additional documents which are translations of literary texts by some of the 8 authors from Dataset 1. A summary of the translations in Dataset 2 can be found in Table 3. All documents were obtained from the Gutenberg project website[1].

Language	**Author**	**Average document length**
English	Charles Dickens	144222
English	Ryder Haggard	97913
French	Alexander Dumas	139681
French	Jules Verne	84124
German	J. W. von Goethe	67671
German	F. Gerstäcker	51655
Spanish	V. Blasco Ibañez	100537
Spanish	B. Perez Galdos	126034

Table 2. Dataset 1 description: 4 languages, 8 authors and 15 documents per author.

Author	**Language (# documents)**
Charles Dickens	French (13)
Alexander Dumas	English(19) Spanish (2)
Jules Verne	English (21) German (3) Spanish (1)
J. W. von Goethe	French (1) English (6)
V. Blasco Ibañez	English (13) French (2)
B. Perez Galdos	English (5)

Table 3. Dataset 2 description: language and number of translated documents.

[1] https://www.gutenberg.org/

3.2 Estimating optimum parameter values

The first parameter to be set is n the number of samples for each sample size i that is necessary to obtain a representative figure for average and dispersion values. An empirical study has been performed with 10 to 2000 samples of each size, using a Random Forest classifier and *leave one out* cross validation. The results of the classification using Fragments and bags-of-words for Dataset 1 are shown on Figure 3.

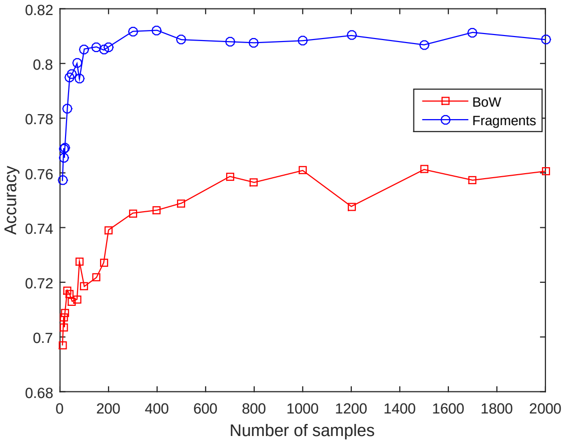


Fig. 3. Number of samples vs. correctly classified documents

The number of correctly classified documents increases as the number of samples increases until a stable value is reached. Fragments and bags-of-words behave differently with more variation in the bags-of-words samples. Two threshold levels can be identified in figure 3, the first threshold is around the value of 200 samples, and the second threshold is around 700 samples where the results are more stable. However, as the computational time is an important factor in text analysis, the selected value for n, the number of samples, is fixed at 200 samples per sample size i.

3.3 Optimum sample size or combination of sample sizes. Number of features.

Once the number of samples is fixed, we need to determine the sample sizes i that will produce the best performing set of features. For each sample size, the proposed method produces a set of 8 features. All sample sizes and their combinations will be empirically tested to evaluate the effect of different numbers of features on the final classification. For this experiment, the following sample sizes (fragments and bags-of-words) have been used: 200, 500, 800, 1000, 1500, 2000, 3000 and 4000.

Combinations of 1, 2, 3, 4, 5, 6, 7 and 8 different sample sizes were taken for both fragment and bag-of-words samples, as well as the combination of both types of samples. In order to optimize the number of features, the combination that produces the highest accuracy with the lowest number of features will be selected.The results, grouped per number of different sample sizes (M) and hence per total number of features, are shown in Figure 4. Figure 4 shows the results for fragments, bags-of-words and the combination of both for Datasets 1 and 2.

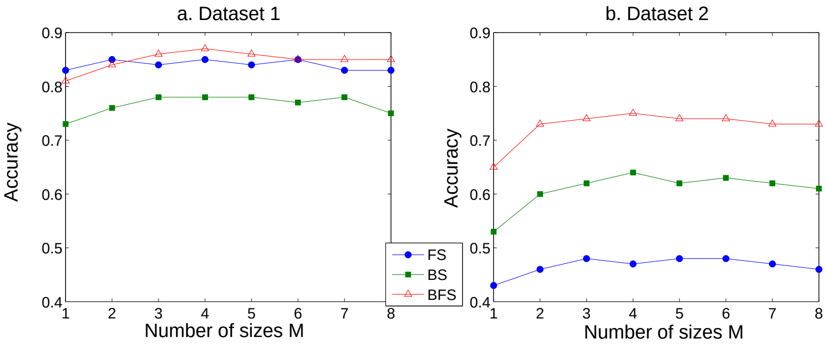


Fig. 4. Accuracy FS, BS and BFS for Datasets 1 and 2

The results from the different combinations of sample sizes show different responses to Dataset 1 and Dataset 2. The different nature of these two datasets explain the different behaviour of the type of samples for each dataset. Fragments are more powerful at discriminating between originals in a balanced setting whereas bags-of-words perform poorly when each author is represented by documents in only one language. On the other hand, bags-of-words provide stronger results for the more difficult problem presented in Dataset 2 where translations are included in the dataset. In both scenarios, the combination of both types of samples, BFS, provides the best results.

In terms of the final size of the feature set, which combines the type of sample and the number of sample sizes i, there is no significant improvement after 2 sizes are combined. The final size of the feature set is therefore $N = 2(8_F + 8_B) = 32$. A closer look at the combination of sizes that produce the best results show sizes 500 and 1000 obtaining the highest accuracy.

Preliminary evaluation of BFS applied to CLAA using the same cross-validation method (*leave one novel out*) and the same dataset as Bogdanova and Lazaridou [3] shows that BFS achieves better classification results (0.47) than high level features without the use of machine translation (0.31). In this particular experiment, 27 documents plus 7 which are translations of one of the 27 are used, with the final dataset being formed by 275 texts extracted from the 34 original documents. For this reason, *leave one novel out* is used to avoid the classifier being trained on texts from the same document (or translations of it). Every time *leave one novel out* is performed on this dataset, a large number of texts are removed from the training data, hence the training set is small, which added to the short length of the texts, affects the overall classification performance. Machine translation methods achieve better results but are limited by the availability of resources in the given languages as well as the requirement to identify the target language beforehand.

4 Conclusion

This paper has presented a feature extraction method for language independent vocabulary richness measurements. Traditional vocabulary richness methods have not performed to state of the art accuracy values in the past and have been replaced with monolingual features such as word n-grams and part-of-speech features. In order to work with multilingual corpora, previous research has used machine translation [3] and lexicons or texts available in several languages such as *wikipedia* [9] or *eurovoc* documents [14]. The proposed method expands traditional vocabulary richness using two types of samples: fragments and bags-of-words of fixed size. It calculates local measurements on those samples as well as the dispersion of those measurements over the samples. The method uses solely deep level intrinsic document measurements and hence no external resources are used.

Our experiments on cross-lingual authorship attribution show that BFS with deep lexical features is suitable for discriminating between authors in multilingual task using a relatively small feature set and no external resources. Even though the accuracy of machine translation based methods is still significantly higher, the experiments reproduced deal with highly popular languages such as English and Spanish, and results for low resource languages are expected to be lower. In these situations, a method based on intrinsic document features such as the one presented in this paper, provides a solution that is not biased by the amount of external resources available. Further work will focus firstly on extensive evaluation of the performance of BFS at a variety of cross-lingual tasks and secondly on the exploration of deep level features used in combination with other language independent methods (implementation-wise) such as character n-grams or methods based on punctuation and sentence length measurements.

References

1. Ahmed Shamsul Arefin, Renato Vimieiro, Carlos Riveros, Hugh Craig, and Pablo Moscato. An Information Theoretic Clustering Approach for Unveiling Authorship

Affinities in Shakespearean Era Plays and Poems. *PLoS ONE*, 9(10):e111445, October 2014.

2. R. Arun, Suresh V. Murty Saradha, R., and C. E. Veni Madhavan. Stopwords and stylometry: a latent Dirichlet allocation approach. In *NIPS workshop on Applications for Topic Models: Text and Beyond*, pages 1–4, 2009.
3. Dasha Bogdanova and Angeliki Lazaridou. Cross-Language Authorship Attribution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, number May, pages 83–86, 2014.
4. John Burrows, David Hoover, David Holmes, Joe Rudman, and Fiona J Tweedie. The State of Non- Traditional Authorship Attribution Studies 2010 : Some Problems and Solutions. *Source*, pages 1–3, 2010.
5. M. A. Covington and J. D. McFall. Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100, 2010.
6. Michael Gamon and Agnes Grey. Linguistic correlates of style : authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics*, 4:611, 2004.
7. V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. *Computational Linguistics*, 3:255–264, 2003.
8. Miroslav Kubát and Jiří Milička. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics*, pages 339–349, 2013.
9. Mari-Sanna Paukkeri, Ilari T. Nieminen, P. Matti, Matti Pöllä, and Timo Honkela. A Language-Independent Approach to Keyphrase Extraction and Evaluation. In *COLING (Posters)*, number August, pages 83–86, 2008.
10. Salvatore Romeo, Dino Ienco, and Andrea Tagarelli. Knowledge-Based Representation for Transductive Multilingual Document Classification. *ECIR 2015*, a:92–103, 2015.
11. Jan Rybicki and Maciej Eder. Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 26(3):315–321, September 2011.
12. Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, March 2009.
13. Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, 26:471–495, 2000.
14. Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. In *Computational Linguistics and Intelligent Text Processing, Third International Conference*, pages 415–424, 2002.
15. Lauren M. Stuart, Saltanat Tazhibayeva, Amy R. Wagoner, and Julia M. Taylor. Style features for authors in two languages. *Proceedings - 2013 IEEE/WIC/ACM International Conference on Web Intelligence*, 1:459–464, 2013.
16. Lawrence Venuti. *The translator's invisibility: A history of translation.* Routledge, 2008.

Profile-based Translation in Multilingual Expertise Retrieval

Hossein Nasr Esfahani, Javid Dadashkarimi, and Azadeh Shakery
{h_nasr,dadashkarimi,shakery}@ut.ac.ir

School of ECE, College of Engineering, University of Tehran, Iran

Abstract. In the current multilingual environment of the web, authors contribute through a variety of languages. Therefor retrieving a number of specialists, who have publications in different languages, in response to a user-specified query is a challenging task. In this paper we try to answer the following questions: (1) How does eliminating the documents of the authors written in languages other than the query language affect the performance of a multilingual expertise retrieval (MLER) system? (2) Are the profiles of the multilingual experts helpful to improve the quality of the document translation task? (3) What constitutes a good profile and how should it be used to improve the quality of translation? In this paper we show that authors' documents are usually related topically in different languages. Interestingly, it has been shown that such multilingual contributions can help us to construct profile-based translation models in order to improve the quality of document translation. We further provide an effective profile-based translation model based on topicality of translations in other publications of the authors. Experimental results on a MLER collection reveal that the proposed method provides significant improvements compared to the baselines.

Keywords: Expert retrieval, multilingual information retrieval, profiles.

1 Introduction

Expert retrieval has achieved growing attention during the past decade. Users in the web aim at retrieving a number of specialists in specific areas [3]. A couple of methods have been introduced for this purpose; retrieving the experts based on their profiles (the candidate-based model), and retrieving the experts based on their published contributions (the document-based model) [3]. The latter approach is usually opted in the literature due to its better performance and its robustness to free parameters [1].

Since there exist a lot of authors who contribute through a variety of languages, using documents written in other languages than the query should intuitively be able to improve the performance of the expertise retrieval system. However scoring documents in such a multilingual environment is challenging. Multilingual information retrieval (MLIR) is a well-known research problem and has been extensively studied in the literature [10]. There are two options for scoring documents written in languages other than the language of the query; translating the query into all the languages of the documents, or representing all the documents in the language of the query. In MLIR it has been shown that the second approach outperforms the first one in the language modeling framework [9]. In the current paper we are going to cast such an approach to multilingual expert retrieval (MLER). Indeed, our new problem is to retrieve experts who are contributing in multiple languages.

In this research we choose the document translation approach for our problem. It is noteworthy that no translated document in the traditional sense is produced, but rather a multilingual representation of the underlying original document that is suitable for retrieval, but not for consumption by a reader, is constructed.

Furthermore, proper weighting of translations has always had a major effect on MLIR performance. Therefore improving the translation model based on user profile can supposedly lead to better MLER performance.

We are trying to answer the following research questions in this paper:

1. How does eliminating the documents of the authors written in languages other than the query language affect the performance of an MLER system?
2. Are the profiles of the multilingual experts helpful to improve the quality of the document translation task?
3. What constitutes a good profile and how should it be used to improve the quality of translation?

Our findings in this paper reveal that multilingual profiles of the experts are useful resources for extraction of expert-centric translation models. To this aim we propose two profile-based translation models using (1) maximum likelihood estimation (PBML), and (2) topicality of the terms (PBT). Indeed translations are chosen based on their contributions in the target language documents of an expert. Our experimental results on a multilingual collection of researchers, specialists, and employees at Tuilberg University [5] reveal that the proposed method achieves better performance on a variety of query topics, particularly in ambiguous ones.

In Section 2 we provide brief history of studies in the literature of MLER and MLIR. In Section 3 the proposed profile-based document translation method is introduced. In Section 4 we provide experimental results of the proposed method and several baselines and then we conclude the paper in Section 5.

2 Previous Work

There have been multiple attempts in the expert finding literature. Most of the research studies aim at retrieving a number of experts in response to a query [4]. Usually a couple of models are employed in an expert retrieval system; candidate-based model and document-based model. Although the former model takes advantage of lower costs in terms of space by providing brief representations for the experts, the latter one achieves better results in some collections [1]. A number of frameworks have been proposed for this aim; model-based frameworks based on statistical language modeling, and frameworks based on topic modeling [2,8]. Balog et al. proposed a language modeling framework in which they first retrieve a number of documents in response to a query and then rank the documents based on their likelihood to the user-specified query. After employing an aggregation module, experts are ranked based on their contributions in the retrieved documents. Theoretically in such a module, there are two factors affecting the retrieval performance; the query likelihood of the documents of the experts, and the prior knowledge about the documents. In the lack of prior knowledge about documents, the documents of an expert are assumed to have uniform distribution. Deng et al. introduced a citation-based model to improve the accuracy of the knowledge about the documents [8]. Nevertheless, the former approach due to its simplicity and its promising results is a popular one in the literature.

In the current multilingual environment of the web, experts are contributing in a variety of languages. In such an environment, a reliable strategy should be employed to bridge the gap between the languages [10,14,7]. A couple of methods for acheiving this goal are proposed; posing either multiple translated queries to the system or retrieving multiple translated documents in response to a query [10]. Although the former method demands an effective rank-aggregation strategy [12], the latter one achieves promising performance in the language modeling framework [10]. These approaches in MLIR can also be adapted to MLER.

3 Profile-based Document Translation

In this section we introduce the proposed expert finding system. The system is going to be used in a multilingual environment to retrieve a number of experts in response to a user-specified query. In this environment the documents of the experts are not necessarily represented in the language of the query.

In MLIR two major approaches are used to overcome this issue. the first approach translates the query into all the languages of the documents and then executes multiple retrieval processes and finally aggregates the results; the second approach represents the documents in all the languages that the query can be posed in and then executes a single retrieval process. Since superiority of the latter approach compared to the former one has been shown in the literature [10], the strategy of the proposed framework lies also on the same road.

To this aim, we use the documents in the profile of an expert to disambiguate translations of terms in the document. Our assumption is that an expert usually publishes articles in one area. So we expect to be able to estimate a robust translation model using the documents of an expert from other languages. In Section 3.1 we delve into the problem by introducing a novel method to build a profile for each expert to improve the translation disambiguation quality, in Section 3.2 we use the proposed profiles to disambiguate translations, and in Section 3.3 we explain the whole expertise retrieval process.

3.1 Building Profiles for Translation Disambiguation

The main goal of the proposed PDT framework is to use local information of the experts' documents to improve the quality of translations. In order to intuitively explain the key idea, consider the following example: suppose an expert has 2 document sets D_1 and D_2 in languages l_1 and l_2 respectively and we want to translate term w_s from one of the documents of D_1 to language l_2. If w_s has two translations w_{t_1} and w_{t_2}, we investigate how these translations are contributing in D_2 documents. The higher the contribution of a translation in D_2, the more likely it is to be the correct translation of w_s. To this end we first construct multiple term distributions in different languages for each expert. We explore two methods to compute the contribution of each term: maximum likelihood and topicality.

Maximum Likelihood Estimation of Contribution of Each Term: In this method we assume that the terms that are more frequent in each expert's documents are more contributing to the whole profile, so we estimate the contribution of each term in a set of documents D as follows:

$$C(w|D) = \frac{\sum_{d\in D} c(w;d)}{\sum_{d\in D} |d|} \tag{1}$$

In Equation 1, $C(w|D)$ indicates the contribution of term w to document set D, $c(w;d)$ indicates the number of occurrences of term w in document d and $N(d)$ is the number of terms in document d.

Topicality Estimation of the Contribution of Each Term: We can use topicality of each term as the measure of contribution of that term to a document set. Zhai Lafferty in [15] proposed an EM based method to compute topicality of terms for pseudo-relevance feedback. We use a similar method: let $\theta_{e_i}^{l_k}$ be the estimated profile model of expert e_i in language l_k based on the relevant document set $D_{e_i}^{l_k} = \{d_1, d_2, .., d_n\}$. According to Zhai & Lafferty we also set λ to some constant to estimate $\theta_{e_i}^{l_k}$. Similar to the model-based PRF we estimate the model with an expectation maximization (EM) method:

$$t^{(n)}(w;l_k) = \frac{(1-\lambda)p_\lambda^{(n)}(w|\theta_{e_i}^{l_k})}{(1-\lambda)p_\lambda^{(n)}(w|\theta_{e_i}^{l_k}) + \lambda p(w|\mathcal{C}^{l_k})} \tag{2}$$

$$p_\lambda^{(n+1)}(w|\theta_{e_i}^{l_k}) = \frac{\sum_{j=1}^{n} c(w;d_j)t^{(n)}(w;l_k)}{\sum_{w'}\sum_{j=1}^{n} c(w';d_j)t^{(n)}(w';l_k)} \tag{3}$$

in which l_k is the k-th language of the expert e_i. λ indicating amount of background noise when generating documents d_j. The obtained language model for expert e_i in language l_k is based on topicality of the words. If a word frequently occurrs in the publications of the expert and also if it is a non-common term through the collection $\mathcal{C}^{l_k}$, it will get a high weight in the profile $\theta_{e_i}^{l_k}$. Our main contribution is to use the language models of the experts in different languages to construct a robust translation model for document translation. Therefore contribution of each term in document set D_{l_k} would be:

$$C(w|D_{e_i}^{l_k}) = p_\lambda(w|\theta_{e_i}^{l_k}) \tag{4}$$

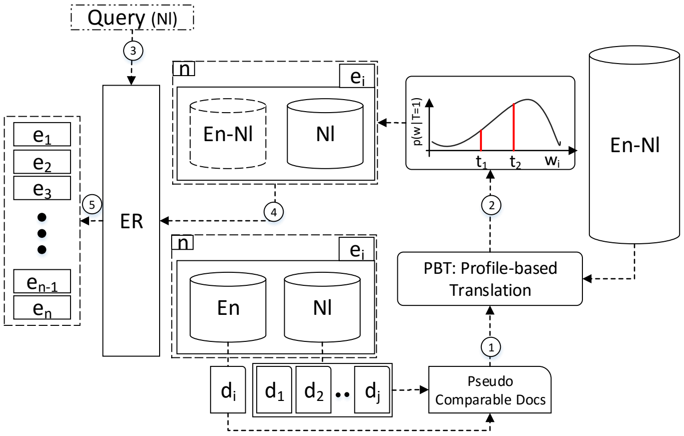


Fig. 1: The proposed expert retrieval framework in multilingual environments.

3.2 Document Translation Based on Cross-lingual Profiles

In this section we introduce the proposed document translation method based on the constructed profiles for each expert. Our goal is to construct translation models for the experts and then to build multilingual documents for them. The translation model for expert e_i is computed as follows:

$$p(w_{t_j}|w_s;e_i) \approx \frac{C(w_{t_j}|D^{l_t}_{e_i})}{\sum_{j'} C(w_{t_{j'}}|D^{l_t}_{e_i})} \tag{5}$$

in which $w_T = \{w_{t_1}, w_{t_2}, .., w_{t_m}\}$ is the set of translation candidates for term w_s from the dictionary. Translations are in language l_t and since we have document translation, w_t is in the source language l_s.

Combining with Other Translation Models: As shown in the cross-lingual information retrieval (CLIR) literature, combining different translation techniques can be useful to obtain a robust translation model [13]. In the proposed framework we also use a general probabilistic dictionary and aim at adapting it to the domain of each expert. We exploit a simple linear interpolation technique:

$$p_\alpha(w_{t_j}|w_s;e_i) = \alpha p(w_{t_j}|w_s;\theta_{par}) + (1-\alpha)p(w_{t_j}|w_s;e_i) \tag{6}$$

where $p(w_{t_j}|w_s;\theta_{par})$ is the translation probability of w_s to w_{t_j} regarding the model obtained from a probabilistic dictionary, and α is a controlling constant.

3.3 The Proposed Expert Retrieval Process

Figure 1 shows the whole process of the proposed expert retrieval system. As shown in the figure, in the first step documents whose languages are different from the query are translated using the PDT framework. This translation technique is based on Rahimi et al. [10] in which all the translations are considered in the retrieval process. Indeed documents are scored based on their relevance to the query. The relevance is computed based on $p_\alpha(w_{t_j}|w_s;e_i)$ obtained in Equation 6. Finally experts are scored based on a document-based model:

$$p(q|e_i) = \sum_d p(q|d)p(d|e_i) \tag{7}$$

For simplicity we estimate $p(d|e_i)$ with a uniform distribution over all the publications of e_i. Moreover we estimate $p(q|d)$ as follow:

$$p(q|d) = \prod_{w \in q} p(w|\theta_d) \tag{8}$$

Similar to [10] we compute $p(w|\theta_d)$ in a multilingual environment as follows:

$$p(w|\theta_d;e_i) = \lambda p_{ml}(w|\theta_d;e_i) + (1-\lambda)p'(w|\mathcal{C}) \tag{9}$$

in which:

$$p'(w|\mathcal{C}) = \frac{\sum_{d\in\mathcal{C}} c_p(w,d)}{N\sum_{d\in\mathcal{C}}|d|}, \quad p_{ml}(w|\theta_d;e_i) = \frac{c_p(w,d)}{N|d|},$$
$$c_p(w,d) = \sum_{u\in d} p(w|u;\theta_{e_i}^{l_k})c(u,d). \tag{10}$$

and N is the number of languages in the collection.

Time Complexity: Although document translation could be time consuming, and profiled based translation exacerbates the problem, but it is worth mentioning that we only translate the terms which are likely to be translated to a query term. Furthermore the EM process is to be computed once per expert and could be done offline, hence this process is totally practical. Nevertheless, the translation model for each expert must be updated when a new document is inserted.

4 Experiments

In this section we provide experimental results of the proposed PDT framework and a number of baselines on a multilingual expert retrieval collection.

4.1 Experimental Setups

We used the bilingual TU expert collection [5] in our experiments. This collection contains a number of documents written by scientists, researchers, and support staff from Tilburg University The collection is provided in an English-Dutch environment. Table 1 shows some statistics one the dataset and Figure 2 shows the contribution of each expert on the set. As shown in Figure 2, experts have enough documents in both languages which makes the dataset suitable for our tests.

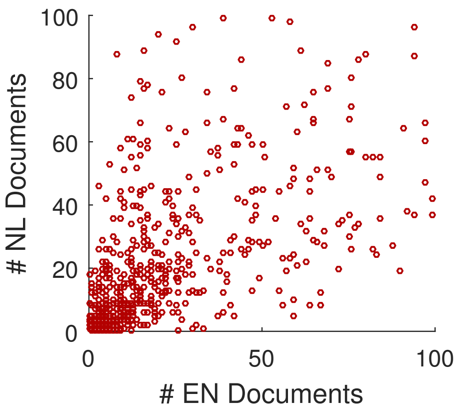


Fig. 2: Distribution of number of documents for each expert.

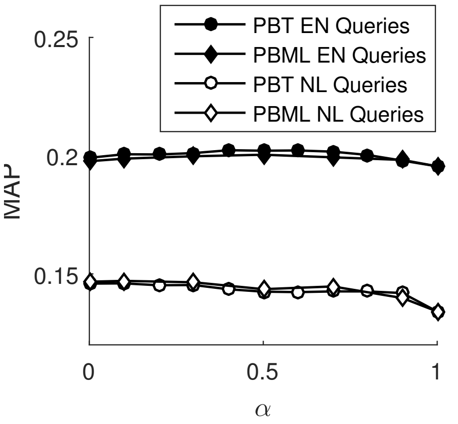


Fig. 3: Sensitivity of the interpolation framework to α

ID	collection	Queries	#queries	#experts	#docs	μ_d	#qrels
TU	Researchers, Scientists, and Employees at Tuilberg University	EN	1,673	893	16,237	1,336	3,936
		NL	2,470	881	20,356	1,204	4,868

Table 1: Collection Statistics. μ_d is the average document length.

Parameter Settings: In all experiments, the Jelinek-Mercer smoothing parameter λ is set to the typical value of 0.9. All free parameters, particularly the constant controlling values of the linear interpolations, are set using 2-fold cross validation over the collection. The noise constant in the EM algorithm is set to 0.7 according to [15].

Evaluation Metrics: We evaluate all the methods based on Mean Average Precision (MAP) of all the retrieved experts as the main evaluation metric. We also report the precision of the top 5 (P@5) and top 10 (P@10) retrieved documents. Statistical differences between the performance of the proposed PDT method and all the baselines are also computed based on two-tailed paired t-test with 95% confidence level on the main evaluation metric [11]. We also provide robustness index (RI) [6] for the last set of our experiments for all the competitive baselines computed as $\frac{N_+ - N_-}{|Q|}$ where $|Q|$ is the number of queries in the collection. N_+ shows the number of queries we have improvements by the proposed method and N_- shows the number of queries in which we have performed worse. Indeed, RI represents the robustness of the method among the query topics.

4.2 Results and Discussions

In this section we report the experimental results of the proposed method and some MLER and CLER baselines. The baselines include MLER based on document translation using (1) top-ranked translations in a probabilistic dictionary

English (EN)						Dutch (NL)					
	TOP-1	MT	PAR	PBT	EN-EN		TOP-1	MT	PAR	PBT	NL-NL
MAP	0.2898	0.2740	0.2898	**0.2911**^{1}	0.2633	MAP	0.2656	0.2458	0.2668	**0.2674**^{012*}	0.2504
P@5	0.1782	0.1637	0.1782	**0.1787**	0.1723	P@5	0.1559	0.1392	0.1568	**0.1571**	0.1474
P@10	0.1208	**0.1244**	0.1208	0.1212	0.1164	P@10	0.1007	0.0981	0.1016	**0.1016**	0.0942

Table 2: Using different translation methods for multilingual expert retrieval. Indicators 0/1/2 denote statistical differences between TOP-1/MT/PAR with confidence of 95%. $*$ shows the confidence is above 90%.

(TOP-1)[1], (2) document translation based on machine translation (MT), (3) weighted translation provided by a probabilistic dictionary (PAR), (4) monolingual retrieval by eliminating documents in out-of-the-context languages (the EN-EN run or the NL-NL one), (5-6) profile-based document translation where profiles are computed w.r.t maximum likelihood (PBML) and topicality (PBT).

Table 2 shows all the results. As shown in the table, all the MLER baselines outperform the simple mono-lingual one. This demonstrates that all the publications of an author, either those in the language of the query or those in other languages, are helpful in our retrieval performance. Although the proposed PBT method outperforms all the baselines in terms of MAP, P@5, and P@10, the improvements in English queries are marginal. The reason for marginal improvements in this dataset goes back to the high performance of the monolingual results. As shown in the table the results of the mono-lingual runs are competitive to the MLER ones (90.45% and 93.64% of PBT in EN-EN and NL-NL runs respectively).

We did further experiments to directly study the effect of the proposed profile-based document translation method. We opted CLER instead of MLER for this purpose. In Table 3 experimental results of a number of CLER runs are provided. These experiments are done only on the documents which are in out-of-the-context languages. To shed light on the effectiveness of the profile-based translation model, we experiment on a subset of the queries which are ambiguous. A query is considered to be ambiguous if at least one of its terms is ambiguous. A term w_t is ambiguous if there exists a term w_s such that $p(w_t|w_s) > 0$ and there exist at least 2 term $w_{t'}$ which $p(w_{t'}|w_s) > \delta$, where δ is a constant value (empirically we set $\delta = 0.2$). As shown in the table, PBT outperforms all the TOP-1, PAR, and PBML baselines in all the evaluation metrics. In the Dutch queries improvements in terms of MAP are also robust (0.2215 out of $[-1, 1]$).

Figure 3 shows the sensitivity of the interpolation framework to α (see Equation 6). As shown in the figure, although the proposed PBT takes advantage of the interpolation approach in both English and Dutch queries, the overall changes are very robust to the parameter. Nevertheless, the results of the PAR baseline without any interpolation with the profile-based translation model drop considerably in Dutch.

[1] We have used a probabilistic dictionary provided by the Google machine translator.

English (EN)					Dutch (NL)				
	TOP-1	PAR	PBML	PBT		TOP-1	PAR	PBML	PBT
MAP	0.1945	0.1955	0.1949	**0.2026**^{012}	MAP	0.1221	0.1341	0.1455	**0.1458**^{01}
P@5	0.1195	0.1208	0.1229	**0.1275**	P@5	0.0712	0.0829	0.0883	**0.093325**
P@10	0.087	0.0867	0.0885	**0.09015**	P@10	0.0585	0.0647	0.0669	**0.0691**
RI	-	-0.1566	-0.0482	**0.0172**	RI	-	0.0196	0.1538	**0.2215**

Table 3: Experimental results for different translation methods for cross-lingual expert retrieval over ambiguous queries.

To sum up our findings we answer the following research questions:

1. How does eliminating the documents of the authors written in languages other than the query language affect the performance of an MLER system? Regarding the competitive mono-lingual results in Table 2 in the TU dataset we can claim that the authors repeat majority of their contributions through languages and so their publications in only one language are almost good but not complete indicators of their expertise. However this kind of conclusion is not valid in real-world data and sometimes authors contribute mainly in a language other than the language of the query.
2. Are the profiles of the multilingual experts helpful to improve the quality of the document translation task? When we want to translate a document of an expert, documents of the expert written in the target language help us to find topical terms. Since correct translations are more likely to be the topical ones we expect to reach a better translation (see Figure 3).
3. What constitutes a good profile and how should it be used to improve the quality of translation? According to Table 3 the proposed PBT method outperforms PBML. This shows that topicality of translations instead of their simple maximum likelihood probabilities are helpful for the document translation task. Further results reveal that interpolating the topical probabilities with values from parallel dictionaries are also useful.

5 Conclusion and Future Work

In this paper we elaborate on the subject of MLER by introducing a novel profile-based document translation method. We have set a number of research questions to this aim and our findings supported the following views: (1) According to our observations, although authors contribute almost similarly in multiple languages, considering all the contributions in different languages can be helpful for expertise retrieval system. Since authors usually repeat their contributions through languages, eliminating documents in out-of-the-context languages does not harm the retrieval performance considerably. (2) Document translation in MLER takes advantage of profile-based translation models. The profile of each expert helps us to opt for topical translations which usually contributes to correct translations. Experimental results on the TU dataset, demonstrate that the proposed profile-based translation approach outperforms a variety of baselines.

An interesting future work of this paper is dynamically learning the interpolation weight between topical probabilities and values from dictionaries based on generality of words. Constructing profiles for a number of expert clusters and employing them in the document translation process will be another future work for this paper.

References

1. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Inf. Ret. pp. 43–50. ACM (2006)
2. Balog, K., Azzopardi, L., de Rijke, M.: A language modeling framework for expert finding. Inf. Proc. & Man. 45(1), 1–19 (2009)
3. Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., Si, L.: Expertise retrieval. Found. Trends Inf. Retr. 6, 127–256 (Feb 2012)
4. Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., Si, L.: Expertise retrieval. Foundations and Trends in If. Ret. 6(2–3), 127–256 (2012)
5. Berendsen, R., Rijke, M., Balog, K., Bogers, T., Bosch, A.: On the assessment of expertise profiles. Journal of the American Society for Inf. Sci. and Tec. 64(10), 2024–2044 (2013)
6. Collins-Thompson, K.: Reducing the risk of query expansion via robust constrained optimization. In: Proceedings of the 18th ACM Conference on Inf. and Know. Manag. pp. 837–846. ACM (2009)
7. Dadashkarimi, J., Shakery, A., Faili, H.: A Probabilistic Translation Method for Dictionary-based Cross-lingual Information Retrieval in Agglutinative Languages. In: Conference of Computational Linguistic (2014)
8. Deng, H., King, I., Lyu, M.R.: Formal models for expert finding on dblp bibliography data. In: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. pp. 163–172. IEEE (2008)
9. Nie, J.Y.: Cross-language information retrieval. Synthesis Lectures on Human Language Technologies 3(1), 1–125 (2010)
10. Rahimi, R., Shakery, A., King, I.: Multilingual information retrieval in the language modeling framework. Inf. Ret. Journal 18(3), 246–281 (2015)
11. Sanderson, M., Zobel, J.: Information retrieval system evaluation: Effort, sensitivity, and reliability. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Res. and Dev. in Inf. Ret. pp. 162–169. ACM (2005)
12. Tabrizi, S.A., Dadashkarimi, J., Dehghani, M., Esfahani, H.N., Shakery, A.: Revisiting optimal rank aggregation: A dynamic programming approach. In: International Conference on the Theo. of Inf. Ret., ICTIR, September (2015)
13. Türe, F., Lin, J.J., Oard, D.W.: Combining statistical translation techniques for cross-language information retrieval. In: COLING. pp. 2685–2702 (2012)
14. Vulic, I., Smet, W.D., Tang, J., Moens, M.: Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. Inf. Process. Manage. 51(1), 111–147 (2015)
15. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the Tenth International Conference on Inf. and Kno. Man. pp. 403–410. ACM (2001)

Extending Automatic Discourse Segmentation for Texts in Spanish to Catalan

Iria da Cunha[1], Eric SanJuan[2], Juan-Manuel Torres-Moreno[2,3], Irene Castellón[4], and Marina Lloberes[4]

[1] Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain
iriad@flog.uned.es
[2] LIA, Université d'Avignon et des Pays de Vaucluse, France
{juan-manuel.torres, eric.sanjuan}@univ-avignon.fr
[3] Ecole Polytechnique de Montréal - Departament de Génie informatique, Montréal, Canada
[4] Universitat de Barcelona - Departament de Lingüística General, Barcelona, Spain
icastellon@ub.edu; mllobesa8@alumnes.ub.edu

Abstract. At present, automatic discourse analysis is a relevant research topic in the field of NLP. However, discourse is one of the phenomena most difficult to process. Although discourse parsers have been already developed for several languages, this tool does not exist for Catalan. In order to implement this kind of parser, the first step is to develop a discourse segmenter. In this article we present the first discourse segmenter for texts in Catalan. This segmenter is based on Rhetorical Structure Theory (RST) for Spanish, and uses lexical and syntactic information to translate rules valid for Spanish into rules for Catalan. We have evaluated the system by using a gold standard corpus including manually segmented texts and results are promising.

Keywords: Discourse Parsing, Discourse Segmentation, Rhetorical Structure Theory, Shallow Parsing, Catalan

1 Introduction

Nowadays discourse parsing is a very prominent research area used in Natural Language Processing (NLP). Recently other NLP applications and approaches that underlie discourse parsing have arose, such as Machine Translation [1], Textual Similarity [2], and Sentiment Analysis and Opinion Mining [3, 4] for example.

In order to develop these applications, discourse segmenters and parsers are needed, as well as discourse level annotated corpora. Several resources have been developed for different languages. Regarding Iberian Peninsula's Romance languages, there are resources for Portuguese [5] and Spanish [2]. However, they have not been developed for Catalan yet. Catalan is a romance language that comes from Latin. It is spoken in several parts of Spain (mainly in Catalonia), Andorra, France (Roussillon) and Italy (Alghero), among others. Despite this number of speakers, this is an under-resourced language and there isn't any discourse annotated corpus available.

Most of the related work mentioned in this paper relies on *Rhetorical Structure Theory* (RST) by [6]. According to this theory, a text can be considered as a hierarchical tree made of elementary discourse units (EDUs) that can work as nucleus or as satellite. While nuclei provide relevant information about the author's point of view, satellites give additional information associated to nuclei. RST discourse parsing is formed by three steps: 1) text segmentation in EDUs, 2) discourse relations analysis and 3) discourse tree building.

There are three strategic approaches to impulse the development of resources for under-resourced languages: a) using the crowd and collaborative platforms; b) using technologies of interoperability with well-developed languages; and c) using Semantic Web technologies and, more specifically, Linked Data. In this work, we focus on the second approach. Therefore, the main goal of this work is to develop and evaluate the first discourse segmenter for Catalan, by adapting the existing discourse segmenter for Spanish (a technologically advanced language).

We use the FreeLing Shallow parser for Catalan [7] as a linguistic resource. The system is formed by linguistic rules based on lexical units (conjunctions and adverbs), discourse markers, syntactic structures and punctuation marks. Furthermore, this work aims to present the annotated corpus developed for evaluating the segmenter. This corpus has been developed as a gold standard open to scientific community.

In section 2, we present the state of the art on discourse segmentation. In section 3, we explain the methodology used to develop our segmenter for Catalan and we go deeper into the system implementation. The details about the experiments and the results are described in section 4. Finally, the conclusions of the current research and the future work are presented in section 5.

2 State of the Art

As stated in [8], discourse is one of the most difficult language levels to process automatically due to its complexity. Actually, this difficulty makes automatic discourse analysis a challenging task because it can be applied to develop several NLP tools. In this sense, [9] and [10] have made surveys upon the current research on discourse parsing and its applications. Some examples are text generation [11–13] and automatic summarization [14–16]. Specifically, research on discourse segmentation has been proved to be useful for different NLP tasks. For example, in [17–19] authors study the relationship between discourse segmentation and compression for sentences in Spanish. They present a method for sentence compression that uses statistical information in order to delete intra-sentence discourse segments, obtaining very good results. This kind of systems has been successfully employed for cinema or TV subtitling, and for elaboration of short messages by mobile companies, among other applications. Also, discourse segmentation has been used for machine translation. For example, [20] use discourse segments to align passages of texts in different languages. Usually, in this field, alignment is done between sentences, but alignment between discourse segments (that is, parts of those sentences) can offer additional information that can be useful for machine translation and other fields.

Existing discourse segmenters employ several strategies; the most productive strategy has been the use of linguistic information (mainly, lexical and syntactic information). The segmenter for English by [21] is based on a statistical model that uses lexical and syntactic features to assign a probability to the insertion of a segment boundary after every word of a sentence. The segmenter for English by [22] is based on linguistic rules (lexical and syntactic) and uses a constituency-based parser. The segmenter for Spanish by [23] is based on linguistic rules adapted to this language and on the grammar of the shallow parser of Freeling.

Another possible strategy to deal with discourse segmentation is machine learning. It would be the case, for example, of the segmenter for French by [24]. However, these types of approaches need a high amount of annotated texts in order to learn and carry out an adequate segmentation. Therefore, nowadays their results are not better than the results of systems based on linguistic information.

Recently, the possibility of developing language-independent segmenters or systems using very few linguistic resources is being explored. It is the case of [25], whose system uses general statistical techniques based on morphological tags, and general linguistic rules. The only language-dependent linguistic resource is a list of discourse markers in the language of the text. The advantages of this type of strategies are that they are easy to implement and require very few resources. Therefore, they are especially adequate for languages without NLP tools. However, currently, the results of these segmenters are not better than the results of the systems designed specifically for particular languages.

In some cases, different strategies have been applied over the same monolingual corpus and have been compared among them, with the aim of determining the most productive strategy for a specific language. It is the case of [26], where both statistical and linguist strategies are applied to segment texts in Basque. Also, constituent and dependency parsing are used. In this work, the strategy based on dependency parsing is the most productive. This is due to the high complexity and particular syntactic characteristics of Basque, since it is an agglutinative language.

3 Methodology

We follow the linguistic strategy to develop a discourse segmenter for Catalan. There are two main reasons for applying this strategy. First, Catalan and Spanish are very similar languages, and the linguistic strategy has been applied for Spanish, obtaining good results. Second, nowadays, in general, this strategy is the most productive for all languages in the state of the art, especially if the used corpus is limited.

3.1 Shallow parsing

The used grammar is an extension of the grammar of the shallow parser for Catalan included in Freeling [7, 27]. The aim of this extension is to detect and re-categorize those words or groups of words that can indicate a boundary between discourse segments in sentences. These rules indicate units that can work as discourse markers. For that, two lexicons of Catalan discourse markers have been processed [28, 29] and 252 markers have been obtained; they were divided in two groups: ambiguous markers (118) and non-ambiguous markers (134). Finally, the 252 rules developed were added to the grammar.

Non-ambiguous markers have been introduced in a new category 'disc-mk' (discourse marker) and therefore have been re-categorized in the grammar of the shallow parser for Catalan. For example: adverbs and adverbial groups (*aleshores*, 'so'; *així doncs*, 'therefore'), prepositional groups (*per causa de*, 'because of') or sequences of lexical units (*tot seguit*, 'next'; *tot i que*, 'although'). On the other hand, the category 'disc-mk-amb' includes composed elements that have been re-categorized from different tags of the shallow parser. For example: *com a mostra* ('as it is shown by') or *després* ('after').

In the case of ambiguous markers, it is necessary to take into account the context where they appear and require advance parsing. For example, the marker *després* ('after') can be just an adverb (see example 1) or a discourse maker (see example 2):

1. *Els resultats mostren que* ***després*** *del test augmentaren els valors.* ('The results show that after the test the values increased.')
2. *Els jugadors de futbol de categoria juvenil van tenir fatiga del sistema nerviós* ***després*** *de realitzar un test de capacitat d'esprints repetits (CER).* ('The football players of the youth category had fatigue of the nervous system after carrying out a repeated sprint test (RST).')

In the grammar, rules related to discourse markers (both ambiguous and non-ambiguous) have priority over the other rules.

3.2 Implementation

The resulting segmenter for Catalan called DiSegCAT is then generated automatically from the segmenter for Spanish.These tools are implemented in Perl, and are based on regular expressions and the Twig XML library. Firstly, DiSeg calls the FreeLing library in order to apply the discourse markers grammar. This first module transforms the syntactic tree generated by the FreeLing Shallow Parser into XML format. The second module applies the rules that detect EDUs boundaries. This module reads all the leaves of the XML syntactic tree. When it recognizes a boundary, the syntactic tree is modified by adding a new node where the boundary is located. This task is iterated twice to find sub-EDUs boundaries inside the EDUs already detected. The third DiSeg module re-reads the new XML syntactic tree to split sentences into coherent EDUs that should contain one or more verbs.

These modules use regular expressions to recognize discourse markers, lemmas and grammatical categories. The idea was the development of a Perl script to translate the code for Spanish discourse segmentation to Catalan. This strategy presuppose that EDUs have the same grammatical structure and segmentation markers are the only elements that are altered. This assumption is not entirely true. There are discourse markers in Spanish that correspond to one or more discourse markers in Catalan. For example, the Spanish marker *para* ('to') is *per* ('to', 'by' or 'through') or *en* ('in') in Catalan. Furthermore, there are grammatical categories tags, such as 'vaux' for auxiliary verbs, that not have a correspondence in the grammar version of Catalan. Therefore, a lexicon that transforms the Spanish DiSeg to the Catalan DiSeg cannot handle these cases but contextual rules can contribute to solve these asymmetries between both languages.

The final implementation is the es2cat.pl perl script that translates the original DiSeg into DiSegCAT. As a consequence, any expansion of the original DiSeg is applied automatically to the Catalan version. This strategy relies on the idea that translating a NLP software is much more easier than translating texts written in natural language. Therefore, the results provided by DiSegCAT are more reliable than the original DiSeg combined with a machine translation system Catalan ↔ Spanish.

4 Experiments and Results

We evaluate system performance over a corpus of manually segmented texts. This type of evaluation is used in previous work in this area [22, 24, 23].

4.1 Corpus

The corpus includes 20 abstracts of research articles in Catalan from the medical domain, extracted from the specialized Journal of Medicine and Physical Activity and Sport *Apunts: Medicina de l'esport*[5]. This journal publishes each article in two languages, Spanish and Catalan, and provides the abstract in these two languages and also in English. This would allow us to perform experiments with parallel corpora in the future. Specifically, for this work, texts published between 2010 and 2013 were selected, in order to have recent documents. Also, texts related to different subjects were selected, such as scoliosis, attention deficit, cardiology, nutrition, etc., in order to guarantee thematic diversity. The textual genre 'abstract' and the medical domain were selected to be able to compare adequately the results of our discourse segmenter for Catalan with the results obtained by the discourse segmenter for Spanish; in the evaluation of this Spanish segmenter, 20 texts with similar characteristics were used.

Once the segmentation criteria were defined for Catalan and the corpus was compiled, manual discourse segmentation was carried out. For that, two annotators were asked to segment each text of the corpus, following the mentioned

[5] http://www.raco.cat/index.php/Apunts/issue/archive

criteria, individually and without questions between them, to avoid biases in the results. Both annotators are linguists and have a wide experience in corpus annotation. After the manual annotation of the 20 texts, both segmentations were compared, in order to determine the inter-annotator agreement. Annotators agreed on 264 discourse segment boundaries and they disagreed on 23 boundaries. Therefore, there was a boundary agreement of 92%.

Following [30] and [31], we have also calculated inter-annotator agreement by using Kappa Cohen in two ways: taking into account words as boundaries and taking into account clauses as boundaries. For the first one, the Kappa value is 0.9556 and, for the second one (that is more conservative), the Kappa value is 0.8674. We consider that these values show that agreement between annotators is high for the task of discourse segmentation.

After this quantitative analysis, a qualitative analysis of the disagreements was done, where we observed that nearly all the disagreements were due to human mistakes. Finally, in the line of work on this topic [23, 26] and [8] indicates, a debate was carried out between annotators to solve disagreements. Thus, agreement was obtained for every case. This final corpus segmented was used as gold standard and will be available online for the scientific community. Table 1 shows the gold standard statistics. As it can be observed in this table, the corpus includes 183 sentences; by contrast, it contains a higher number of discourse segments (280), which means that intra-sentence discourse segmentation is productive.

	Total	**Longuest text**	**Shortest text**	**Average**
Num. of words	4 676	317	91	233.80
Num. of sentences	183	17	4	9.15
Num. of segments	280	24	8	14.00

Table 1. Gold Standard statistics.

4.2 Evaluation

We consider two baseline segmenters to compare our results:

- Baseline$_1$: it inserts boundaries before coordinating conjunctions.
- Baseline$_2$: it considers all complete sentences like discourse segments. This baseline will have a precision of 100%, because all detected segments will be correct, since sentences are considered discourse segments.

The results obtained are shown in Table 2. Differences between DiSegCAT and Baseline$_2$ scores are all significant based on 12-fold t-test with p-value < 0.05.

Since there is not another discourse segmenter for Catalan we cannot compare our results with another system. This is a current situation when working on

System	F-Score	Precision	Recall
DiSegCAT	**75%**	68%	**85%**
Baseline$_1$	52%	44%	65%
Baseline$_2$	18%	**100%**	10%

Table 2. Results of our experiment.

local languages. We did try to combine DiSeg for Spanish with mainstream available translators form Spanish to Catalan. But translation is an even more complex problem than discourse analysis and in the case of our corpus in Spanish, translators were not efficient with high rate of errors including over short multiword expressions.

However, we find that our results are similar to those obtained for other languages by using similar strategies of segmentation, such as for English (F-Score = 83%) and Spanish (F-Score = 80%).

As it can be observed in Table 2, our system (DiSegCAT) obtains the best F-Score (75%), in comparison with Baseline$_1$ (52%) and Baseline$_2$ (18%). As expected, Baseline$_2$ obtains 100% of precision, since it considers sentences as segments; however, it obtains only 10% of recall, since it does not detect intra-sentence segments. Baseline$_1$ has 44% of precision and 65% of recall. Although this baseline detects correctly some segments (because coordinated clauses can be also discourse segments), both results (precision and recall) are worst than the results obtained by DiSegCAT. These results mean that our algorithm outperforms baselines including linguistic information.

Regarding the results obtained by DiSegCAT, its performance is better for recall than for precision (85% and 68%, respectively). After obtaining these quantitative results, we have carried out a qualitative analysis in order to find different types of errors of the system.

With respect to precision, the main error is related to coordination. See for example the following segments, obtained automatically by DiSegCAT[6]:

```
[El nostre objectiu fou establir quins paràmetres antropomètrics]
[i de maduració es correlacionen amb el rendiment
en rem-ergòmetre en una mostra de 114 adolescents d'ambdós sexes,
sense experiència prèvia en rem.]
```

```
[We aimed to establish which anthropometric]
[and maturity offset parameters correlate with
rowing ergometer performance in a sample of 114 adolescent,
rowing-inexperienced boys and girls.]
```

Here, following our segmentation criteria and rules, the correct segmentation would be:

[6] English translation of examples has been extracted from the papers published by the authors in the journal *Apunts: Medicina de l'esport*.

[El nostre objectiu fou establir quins paràmetres antropomètrics i de maduració es correlacionen amb el rendiment en rem-ergòmetre en una mostra de 114 adolescents d'ambdós sexes, sense experiència prèvia en rem.]

The passage *quins [...] rem* is the direct object of the main verb (*fou establir*, "was to establish") of the sentence. Therefore, it should not be segmented. Nevertheless, this direct object includes a coordination that contains the conjunction *i* ("and") and the finite verb *es* ("is"). Thus, the system segments after the conjunction, since one of the rules indicates that a passage written after this conjunction should be a discourse segment if it includes a finite verb. As it can be observed, in this case, the performance of this rule is not adequate. We should find a solution for this problem in the future. Coordination is also one of the main difficulties found in the performance of the discourse segmenter for Spanish [32].

With regard to recall, the main problem is related to segments that are not explicitly marked in the text. See for example the following segment, obtained automatically by the system:

Té un cost baix, és massiva i de fàcil aplicació.

It is low cost, it is massive and easy to use.

The adequate segmentation of this passage should be:

[Té un cost baix]
[és massiva i de fàcil aplicació.]

The second segment includes a finite verb, but there is no specific mark indicating that it is a discourse segment (the comma is not included in our rules as a boundary mark, since it would over-generate discourse segments). In the future, we plan to study strategies to solve this limitation, although it is a difficult issue.

5 Conclusions and Future Work

This paper presents the first discourse segmentation system for Catalan based on RST. The segmenter uses simple linguistic rules and promising results have been obtained in our experiments. This system could be used in different tasks in the context of NLP. For example, segmentation of (too) long sentences which are difficult to parse; elimination of text segments in Sentence Compression and Automatic Text Summarization; Alignment of text in different languages for Machine Translation, etc. Also, the segmenter can be the basis for further development of an automatic discourse parsing system for Catalan, since this tool does not exist so far.

The methodology used to develop this segmenter for Catalan is very similar to that used for Spanish. In fact, the good results obtained show us that this methodology is probably valid in general for Romance languages. The future experiments will be conducted on this line. We also plan to expand the size of the corpus, adding texts of other genres (such as news reports) and other domains (such as Linguistics or Economy).

Acknowledgements

This work has been partially supported by a Ramón y Cajal research contract (RYC-2014-16935) and the research project APLE 2 (FFI2009-12188-C05-01) of the Institute for Applied Linguistics (IULA).

References

1. Guzmán, F., Joty, S., Márquez, L., Nakov, P.: Using discourse structure improves machine translation evaluation. In: ACL 2014. Volume 1., Association for Computational Linguistics (ACL) (2014) 687–698
2. da Cunha, I., Vivaldi, J., Torres-Moreno, J.M., Sierra, G.: SIMTEX: An Approach for Detecting and Measuring Textual Similarity based on Discourse and Semantics. Computación y Sistemas **18**(3) (2014) 505–516
3. Trnavac, R., Taboada, M.: Discourse structure and attitudinal valence of opinion words in sentiment extraction. In: 88th Meeting of the Linguistic Society of America, Poster, Minneapolis, MN (2014)
4. Chenlo, J., Hogenboom, A., Losada, D.: Sentiment-based Ranking of Blog Posts using Rhetorical Structure Theory. In: NLDB'13, Salford, UK (2013) 13–24
5. Pardo, T., Nunes, M.: On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. Journal of Theoretical and Applied Computing **15**(2) (2008) 43–64
6. Mann, W., Thompson, S.: Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse **8**(3) (1988) 243–281
7. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M.: FreeLing 1.3. Syntactic and semantic services in an open-source NLP library. In: LREC'06, Genoa, Italy, ELRA (2006) 48–55
8. Hovy, E.: Annotation. In: Tutorial Abstracts of ACL 2010, Uppsala, Sweden, ACL (July 2010) 4
9. Taboada, M., Mann, W.: Applications of Rhetorical Structure Theory. Discourse Studies **8**(4) (2006) 567–588
10. da Cunha, I., Torres-Moreno, J.M., Sierra, G.: Aplicaciones lingüísticas del análisis discursivo automático. In Ruiz, L.; Álvarez, M.R., ed.: Comunicación Social en el Siglo XXI Vol. II, Centro de Lingüística Aplicada (2011) 919–923
11. Hovy, E.: Automated discourse generation using discourse structure relations. Artificial Intelligence **63** (1993) 341–385
12. Dale, R., Hovy, E., Rösner, D., O., S.: Aspects of Automated Natural Language Generation. Springer (1992)
13. O'Donnell, M., Mellish, C., Oberlander, J., Knott, A.: ILEX: An architecture for a dynamic Hypertext generation system. Natural Language Engineering **7** (2001) 225–250

14. Marcu, D.: The Theory and Practice of Discourse Parsing Summarization. Institute of Technology, Massachusetts (2000)
15. Radev, D.: A common theory of information fusion from multiple text sources. Step one: Cross document structure. In: 1st SIGdial Workshop on Discourse and Dialogue, Hong-Kong (2000) 74–83
16. Pardo, T., Rino, M.: DMSumm: Review and assessment. In: PorTAL'02, Faro, Portugal:Springer (2002) 263–274
17. Molina, A., Torres-Moreno, SanJuan, E., da Cunha, I., Sierra, G.J.M., Velázquez-Morales, P.: Discourse Segmentation for Sentence Compression. In: LNAI 7094. Volume abs/1212.3493., Berlin: Springer (2011) 316–327
18. Molina, A., Torres-Moreno, J.M., da Cunha, I., SanJuan, E., Sierra, G.: Sentence Compression in Spanish driven by Discourse Segmentation and Language Models. CoRR **abs/1212.3493** (2012)
19. Molina, A., Torres-Moreno, J.M., SanJuan, E., da Cunha, I., Sierra, G.: Discursive Sentence Compression. In: LNCS. Volume 7817., Springer (2013) 394–407
20. Ghorbel, H., Ballim, A., Coray, G.: ROSETTA: Rhetorical and Semantic Environment for Text Alignment. In Rason, P., e.a., ed.: Proceedings of Corpus Linguistics 2001. (2001) 224–233
21. Soricut, R., Marcu, D.: Sentence Level Discourse Parsing Using Syntactic and Lexical Information. In: Conference of the North American Chapter of the ACL on Human Language Technology, Edmonton, Canada (2003) 149–156
22. Tofiloski, M., Brooke, J., Taboada, M.: A Syntactic and Lexical-based Discourse Segmenter. In: 47th Meeting of the ACL, Singapur (2009) 77–80
23. da Cunha, I., SanJuan, E., Torres-Moreno, J.M., Lloberes, M., Castellón, I.: DiSeg: Un segmentador discursivo automático para el español. Procesamiento del Lenguaje Natural **45** (2010) 145–152
24. Afantenos, S., Denis, P., Muller, P., Danlos, L.: Learning recursive segments for discourse parsing. In: LREC'10, Malta, ELRA (2010) 3578–3584
25. Saksik, R., Molina, A., Carneiro Linhares, A., Torres-Moreno, J.M.: Segmentacao discursiva automática: uma avaliacão preliminar em francés. In: 4th Workshop RST and Discourse Studies, STIL'13, Fortaleza, Brasil (2013) 40–49
26. Iruskieta, M., Aranzabe, M., Díaz de Ilarraza, A., Gonzalez, I., Lersundi, M., Lopez de Lacalle, O.: The RST Basque TreeBank: an online search interface to check rhetorical relations. In: 4th Workshop RST and Discourse Studies, STIL'13, Fortaleza, Brasil (2013) 40–49
27. Padró, L., Collado, M., Reese, S., Lloberes, M., Castelló, I.: Freeling 2.1: Five years of open-source language processing tools. In: LREC'10, Malta (2010) 931–936
28. Alonso, L.: Representing Discourse for Automatic Text Summarization via Shallow. PhD thesis, Univ. de Barcelona (2005)
29. Generalitat-Catalunya: Redacció de documents: mots connectors. Barcelona: Departament de Justícia. (2014)
30. Iruskieta, M.: The relational discourse structure in pragmatics: description and evaluation in Computational Linguistics. PhD thesis, University of the Basque Country (2014)
31. Iruskieta, M., Díaz de Ilarraza, A., Lersundi, M.: Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. Corpus Linguistics and Linguistic Theory (CLLT) **9**(1) (2013) 1–32
32. da Cunha, I., SanJuan, E., Torres-Moreno, J.M., Cabré, M., Sierra, G.: A Symbolic Approach for Automatic Detection of Nuclearity and Rhetorical Relations among Intra-sentence Discourse Segments in Spanish. In: LNCS. Volume 7181., Springer (2012) 462–474

A New Image Analysis Framework for Latin and Italian Language Discrimination

Darko Brodić[1], Alessia Amelio[2], and Zoran N. Milivojević[3]

[1] University of Belgrade, Technical Faculty in Bor, V.J. 12, 19210 Bor, Serbia
[2] DIMES University of Calabria, Via P. Bucci Cube 44, 87036 Rende (CS), Italy
[3] College of Applied Technical Sciences, Aleksandra Medvedeva 20, 18000 Niš, Serbia
dbrodic@tf.bor.ac.rs, aamelio@dimes.unical.it,
zoran.milivojevic@vtsnis.edu.rs

Abstract. The paper presents a new framework for discrimination of Latin and Italian languages. The first phase maps the text in the given language into a uniformly coded text. It is based on the position of each letter of the script in the text line and its height, derived from its energy profile. The second phase extracts run-length texture measures from the coded text given as 1-D image, by producing a feature vector of 11 values. The obtained feature vectors are adopted for language discrimination by using a clustering algorithm. As a result, the distinction between the two languages is perfectly realized with an accuracy of 100% on a complex database of documents in Latin and Italian languages.

Keywords: Clustering, Document analysis, Image processing, Information retrieval, Italian language, Statistical analysis

1 Introduction

Information retrieval represents one of the areas of natural language processing. It finds the objects, which usually represent documents of an unstructured nature (usually text) that satisfy an information need from within large collections [11]. Typically, the vector space model is used for similarity distinction between the documents. However, the cross-language information retrieval is still a challenge. It is especially expressed between very similar languages or languages that evolved one from another.

The Latin language was originally spoken in the region around Rome called Latium. As a consequence of Roman conquests, Latin was quickly spread over a larger part of Italy and wider. Accordingly, it has begun the formal language of the Roman Empire. After its collapse, Latin language evolved into the various Romance languages. However, it was still used for writing. Furthermore, the Latin language was a lingua franca, which was used for scientific and political affairs, for more than a thousand years. Up to now, ecclesiastical Latin language has remained the formal language of the Roman Catholic Church. As a consequence, it is the official language of the Vatican. Although Latin language is not a live language, it is not a dead language. It is still partly in use.

Today, the Latin language is usually taught in order to translate Latin texts into modern languages. Because of this long tradition and of the influence on the modern languages, the study of Latin is extremely important for linguistic research. Italian language is one of the languages from the Romance language group, which is the closest to the Latin language. It comprises many dialects from the North to the South of Italy. However, the standard Italian language is virtually the only written language. Today, the standard Italian language is virtually the only dialect of culture in modern Italy, which is used as the language of intercommunication between different parts of Italy. To the very best of the author's knowledge, some aspects of evolving Latin into modern Italian language have been researched. Still, these aspects were completely linguistics in nature [5]. In contrast, we conducted the research in the direction of safe automatic differentiation of these languages in unsupervised manner.

In this paper, we propose a novel framework for the distinction between languages that evolved one from another. As an example, we use Latin and modern Italian languages. The framework includes the following stages: script coding, run-length texture analysis and clustering. The main novelty of the framework is the extension of a state-of-the-art clustering method and its application on document features for discrimination of languages evolved one into another. Because we deal with discrimination problem, unsupervised method is appropriate. The distinction between the two related languages is perfectly realized with an accuracy of 100%, which outperforms competitor methods.

The paper is organized in the following manner. Section 2 describes the proposed framework. Section 3 explains the experiment. Section 4 gives the results of the experiment and discusses them. Section 5 makes a conclusion.

2 The Proposed Framework

Our framework for Latin and modern Italian language discrimination is composed of the following three steps: (i) script coding, (ii) texture analysis, (iii) clustering. Script coding adopts the approach previously introduced by Brodić et al. [4]. In fact, it demonstrated to be successful for solving a critical task of closely related language discrimination [3]. In particular, given the text document as input, it maps each letter of the document to only four codes based on the corresponding position in the text line, representing the gray-level pixels of a 1-D image. Then, texture analysis is performed on the produced image in order to extract run-length texture features. In order to select the feature representation, three well-known types of texture features, run-length, co-occurrence and ALBP, have been evaluated on benchmark datasets of the same languages. Results demonstrated that run-length features obtain the best performances in language discrimination in this context. These features are discriminated by a new clustering method in order to detect classes representing documents written in two different languages.

2.1 Script Coding

Text documents can be divided into text lines. Furthermore, each text line can be segmented by considering the energy of the script signs [9] into the four virtual lines [20]: top-line, upper-line, base-line and bottom-line. These lines track the following vertical zones in the text line area [20]: upper zone, middle zone and lower zone. The letters can be categorized based on their position in vertical zones of the line, that represents their energy profile. The short letters (S) are located into the middle zone only. The ascender letters (A) occupy the middle and upper zones. The descendent letters (D) are spread into the middle and lower zones. The full letters (F) enlarge over all vertical zones. Consequently, all letters can be classified as belonging to four different script types [4]. Fig. 1 depicts the script characteristics according to their position in the baseline.

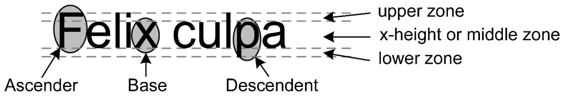


Fig. 1. Virtual lines and vertical zones in the text line.

Each script type can be mapped into a different number code. Because there are only four script types, mapping is performed to four number codes {0, 1, 2, 3}. Then, these codes are associated with four different gray levels to create an image. Fig. 2 illustrates the correspondence between script type number codes and gray levels.

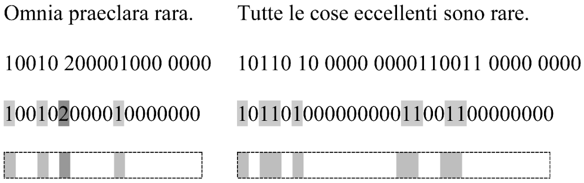


Fig. 2. Script type number codes and their corresponding gray levels of the 1-D image.

Consequently, each text document is translated into a set of number codes {0, 1, 2, 3} corresponding to pixels of only four gray levels. It obtains a textured 1-D image I, which can be analyzed by adopting the texture analysis.

2.2 Texture Analysis

Texture quantifies the intensity variation in the image area [16]. Hence, it is a powerful tool for the extraction of important properties like image smoothness, coarseness and regularity. Accordingly, the texture is useful to compute image statistical measures. Run-length statistical analysis is adopted to retrieve texture features and to evaluate texture coarseness [8]. A run is a set of consecutive pixels with the same gray-level value in the specific texture direction. The fine textures are characterized by long runs, while coarse textures include short runs.

Let I be an image of X rows, Y columns and L gray levels. The first step consists in building the run-length matrix $\mathbf{P}$. It is created by fixing a direction and then counting how many runs are encountered for each gray level and length in that direction. Accordingly, a set of consecutive pixels with identical intensity values identifies a gray-level run. The row number of $\mathbf{P}$ is equal to L, i.e. the number of gray levels, while the column number of $\mathbf{P}$ is equal to the maximum run length R. In our case, a single element of the run-length matrix $P(i, j)$ at position (i, j) represents the number of times a run of gray-level i and of length j occurs inside the image I (in our case, 1-D image).

Different texture features can be extracted from the $\mathbf{P}$ matrix [8]: (i) Short run emphasis (SRE), (ii) Long run emphasis (LRE), (iii) Gray-level non-uniformity (GLN), (iv) Run length non-uniformity (RLN), and (v) Run percentage (RP). The extraction of texture features from $\mathbf{P}$ includes also the following two measures [6]: (i) Low gray-level run emphasis (LGRE) and (ii) High gray-level run emphasis (HGRE). In Dasarathy et al. [7], other four texture features are proposed, based on the joint statistical measure of gray level and run length. They are: (i) Short run low gray-level emphasis (SRLGE), (ii) Short run high gray-level emphasis (SRHGE), (iii) Long run Low gray-level emphasis (LRLGE), and (iv) Long run high gray-level emphasis (LRHGE).

In this way, run-length statistical analysis extracts a total of 11 feature measures, defining a 11-dimensional feature vector for language representation.

2.3 Clustering

The aforementioned run-length feature vectors, each representing a document in Latin or modern Italian languages, are subjected to unsupervised classification by a clustering technique. It is adopted for discriminating between documents written in Latin language and documents written in modern Italian language. In order to find the classes in the data, we adopt the Genetic Algorithms Image Clustering for Document Analysis algorithm (GA-ICDA), previously introduced by Brodić et al. [3], modified to be suitable for languages evolved one into another. We call the modified version of this algorithm *Genetic Algorithms Image Clustering for Document Analysis-Plus* (GA-ICDA$^+$). Next, we recall the main concepts underlying GA-ICDA and propose the modifications for GA-ICDA$^+$.

GA-ICDA is a bottom-up clustering method representing the set of documents written in different languages or scripts as a weighted graph $G = (V, E, W)$. Each node $v_i \in V$ is a document and each link $e_{ij} \in E$ connects two nodes v_i and v_j to each other. A weight $w_{ij} \in W$ associated to the link e_{ij} represents the similarity among the nodes v_i and v_j. For each node v_i, only a set of the other nodes $V \setminus v_i$ in G is considered. This set is called h-nearest neighborhood of v_i [1]. It represents the set of nodes whose corresponding documents are the most similar to the document associated to v_i. Similarity between two nodes v_i and v_j is calculated as:

$$w_{ij} = e^{-\frac{d(i,j)^2}{a^2}}, \tag{1}$$

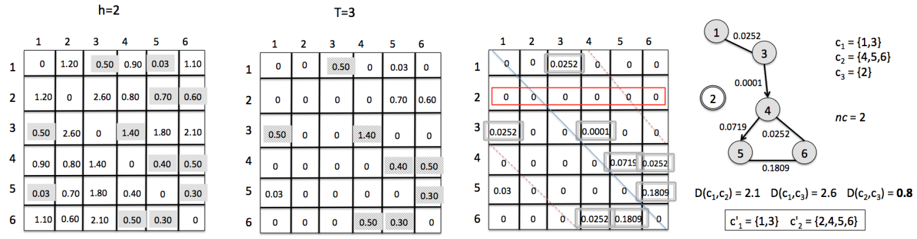
where a is a scale parameter and $d(i, j)$ is the distance between the document feature vectors of v_i and v_j. The L_1 norm is adopted as distance, while h is a parameter influencing the size of the neighborhood [1]. The h-nearest neighbor nodes of v_i are denoted as $nn^h_{v_i} = \{nn^h_{v_i}(1), ..., nn^h_{v_i}(k)\}$, where k is the number of h-nearest neighbors. Then, a mapping f is defined between each node in V and an integer label, $f : V \rightarrow \{1, 2, .., n\}$ $n = |V|$, realizing a node ordering. Finally, the difference is calculated between the label corresponding to the node $f(v_i)$ and the labels corresponding to the nodes in $nn^h_{v_i}$, $|f(v_i) - f(nn^h_{v_i}(j))|$ $j = 1...k$. Each node v_i in G is connected only to the nodes in $nn^h_{v_i}$ whose label difference is less than a given threshold value T. It implies that only similar and "spatially" close nodes are connected to each other in G. The obtained node connections, weighted by the similarity values, are represented in terms of the adjacency matrix $\mathbf{M}$ of G. Then, G is subjected to a genetic method for finding the connected components representing the clusters of documents. After that, for correcting the local optima, a merging procedure is applied on the found clusters. In particular, pairs of clusters having minimum mutual distance are selected and repeatedly merged, until a fixed cluster number is reached. The distance is computed as the L_1 norm between the two farthest document feature vectors, one for each cluster.

The first introduced modification in GA-ICDA$^+$ is the similarity computation among the graph nodes. The inner complex and variegate structure of the evolved language, like modern Italian, determines naturally higher distance values computed between the document feature vectors. Such a phenomenon may cause an anomaly in the similarity computation in Eq. (1). Consider v_i as a node in G with associated document feature vector d_i. If the distance $d(i, j)$ between the vectors d_i and d_j of the nodes v_i and v_j is particularly high, because of the power by 2, the numerator of the exponent $\frac{d(i,j)^2}{a^2}$ is very high, determining a similarity value which is zero. If it occurs much often for different pairs of document feature vectors, the adjacency matrix $\mathbf{M}$ corresponding to the similarity matrix will be unjustifiably very sparse. In order to overcome this problem, the exponent of $d(i, j)$ in Eq. (1) which is currently 2, is substituted by a parameter α for obtaining a more flexible and smoothed characterization of the similarity. Consequently, w_{ij} in Eq. (1) begins:

$$w_{ij} = e^{-\frac{d(i,j)^{\alpha}}{a^2}}. \tag{2}$$

The second introduced modification is the graph construction. Specifically, consider the second step of the procedure where, for each node v_i, only the h-nearest neighbors are maintained, which are "spatially" close to v_i, given a node ordering f. It is clear that it determines a reduction in the number of neighbors, and consequently in the number of outgoing links, for each node v_i. It obtains in most cases a better characterization of the graph connected components. When the document graph is particularly complex, like in this task of capturing differences between languages evolved one into another, a low value of the threshold T is necessary for determining good components. However, it causes the presence of isolated nodes, for which all the nearest neighbors are removed by the threshold T. In GA-ICDA this situation is not considered, because we obtain good components even if the T value is higher. Here we relax this constraint, by managing the presence of isolated nodes. They are "singleton" nodes for the genetic procedure, which is not able to add them inside any connected component, because of the absence of node neighbors. At the end of the procedure, they will be considered as "singleton" clusters and automatically managed by the final bottom-up strategy.

Fig. 3 shows an example of GA-ICDA$^+$ execution. From left to right, for each node in the distance matrix (6 nodes), the algorithm finds the 2-nearest neighbors (in grey). Then, for each node, the algorithm finds the neighbors with label difference smaller than $T = 3$ with respect to the label of that node (in dotted grey), making the node 2 isolated. The adjacency matrix is obtained by computing the similarity values from the distance values by adopting Eq. (2) ($\alpha = 1.5$). c_1, c_2 and c_3 are the clusters detected from the genetic algorithm. c'_1 and c'_2 are the final clusters detected from the bottom-up merging procedure, with fixed cluster number $nc = 2$. They are obtained by computing the distances of cluster pairs and merging the singleton cluster c_2 with c_3 exhibiting the minimum distance value of 0.8.



h=2

	1	2	3	4	5	6
1	0	1.20	0.50	0.90	0.03	1.10
2	1.20	0	2.60	0.80	0.70	0.60
3	0.50	2.60	0	1.40	1.80	2.10
4	0.90	0.80	1.40	0	0.40	0.50
5	0.03	0.70	1.80	0.40	0	0.30
6	1.10	0.60	2.10	0.50	0.30	0

T=3

	1	2	3	4	5	6
1	0	0	0.50	0	0.03	0
2	0	0	0	0	0.70	0.60
3	0.50	0	0	1.40	0	0
4	0	0	0	0	0.40	0.50
5	0.03	0	0	0	0	0.30
6	0	0	0	0.50	0.30	0

	1	2	3	4	5	6
1	0	0	0.0252	0	0	0
2	0	0	0	0	0	0
3	0.0252	0	0	0.0001	0	0
4	0	0	0	0	0.0719	0.0252
5	0.03	0	0	0	0	0.1809
6	0	0	0	0.0252	0.1809	0

Fig. 3. Example of GA-ICDA$^+$ execution.

3 Experimentation

As example of framework usage, an experiment is performed on a complex custom oriented database, publicly available at [10], composed of a set of 90 documents in Latin and modern Italian languages. Specifically, 50 out of 90 documents are given in Latin language and 40 out of 90 documents are given in modern Italian language. Documents count from 400 to 6000 characters each. 40 out of 50 Latin documents are extracted from Cicero's works (106 BC - 43 BC), in particular from *De Inventione*, *De Oratore*, *De Optimum Genere Oratorum*, *De Natura Deorum* and *De Officiis.* 10 out of 50 Latin documents are extracted from Virgil's *Aenead* (70 BC - 19 BC). The documents from the two different authors belong to a different historical period and the writing style of the two authors is also different. Consequently, recognition of common language is difficult. Modern italian documents are extracted from two well-known Italian newspapers, *Il Sole 24Ore* and *La Repubblica*, and from websites. In particular, 20 out of 40 modern Italian documents are excerpts from newspapers and 20 out of 40 modern Italian documents are excerpts from the web. The writing style of the newspapers excerpts is different, because more "technical", than the writing style of the excerpts from the web, which is more "linear".

4 Results and Discussion

Next, we demonstrate the efficacy of our framework as a combination of feature representation and clustering method, in correctly discriminating between Latin and modern Italian documents. Specifically, we show in Table 1 the clustering results obtained from our framework (named as GA-ICDA$^+$) on the custom oriented document database and compare them with the clustering results obtained from other five algorithms on the same database. They are three clustering methods, Hierarchical Clustering, K-Medians and Self-Organizing-Map (SOM), which are different well-known strategies for text document categorization [13],[15],[19]. In particular, we chose to adopt K-Medians instead of K-Means because the first one uses the same L_1 norm as our method GA-ICDA$^+$ and because it is more robust to outliers than K-Means. The other two algorithms are the GA-IC framework for image database clustering [1] and the GA-ICDA framework [3], which is the extension of GA-IC for document database clustering, without the modifications introduced for GA-ICDA$^+$. All the algorithms, K-Medians, hierarchical clustering, SOM, GA-IC and GA-ICDA adopt the same run-length feature vector representation used from GA-ICDA$^+$.

Clustering results are showed in terms of five methods for performance evaluation: precision, recall and f-measure indexes [2],[12], purity, entropy, Normalized Mutual Information (NMI) [2],[17],[18] and Adjusted Rand Index (ARI) [14]. Precision, recall and f-measure are reported separately for each language class (Latin and modern Italian) in correspondence to each algorithm. For the other performance measures, purity, entropy, NMI and ARI, a single overall value is reported for each algorithm. Purity, entropy, NMI and ARI are well-known performance measures for clustering evaluation. On the contrary, the computation of precision, recall and f-measure requires that the correspondence between each cluster detected from the algorithm and the true language class is known. Consequently, we associate each cluster with the true language class whose corresponding number of documents is in majority in that cluster. The number of clusters nc found from the algorithms is also reported.

A trial and error procedure has been adopted on benchmark documents, different from the documents in the considered database, for tuning the algorithms parameters. The parameter values providing the best possible results on the benchmark documents have been adopted for clustering the custom oriented document database. Consequently, in K-Medians algorithm, the number of clusters is fixed to 2. In SOM algorithm, the dimension of a neuron layer is 1×2. The number of training steps for initial covering of the input space is 100 and the initial neighborhood size is 3. The distance between two neurons is computed as the number of steps separating each other. Hierarchical clustering adopts a bottom-up agglomerative strategy using L_1 norm for distance computation. Average linkage is used for cluster distance evaluation. The obtained dendrogram is "horizontally" cut to obtain a number of clusters which is equal to 2. The h value of the neighborhood is fixed to 33 for GA-IC and GA-ICDA and to 43 for GA-ICDA$^+$ and the T threshold value to 9 for GA-ICDA and to 7 for GA-

ICDA$^+$. The α parameter for the similarity computation in GA-ICDA$^+$ is fixed to 1.5.

The algorithms have been implemented in MATLAB R2012a. Experiments have been run on a Desktop computer quad core 2.3GHz 4GB RAM and Windows 7. Each algorithm has been executed 100 times and the average values of each performance measure together with the standard deviation values (in parenthesis) have been reported. Our framework takes 55 s for each execution on the database of 90 documents.

Table 1. Results of Latin and modern Italian document clustering.

	classes	Precision	Recall	F-Measure	Purity	Entropy	NMI	ARI	nc
GA-ICDA$^+$	Latin	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	0.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	2
	modern Italian	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)					
GA-ICDA	Latin	1.0000 (0.0000)	0.9000 (0.0000)	0.9474 (0.0000)	0.9444 (0.0000)	0.2237 (0.0000)	0.7428 (0.0000)	0.7878 (0.0000)	2
	modern Italian	0.8889 (0.0000)	1.0000 (0.0000)	0.9412 (0.0000)					
GA-IC	Latin	0.8113 (0.0000)	0.8600 (0.0000)	0.8350 (0.0000)	0.8111 (0.0000)	0.6215 (0.0000)	0.2967 (0.0000)	0.3803 (0.0000)	2
	modern Italian	0.8108 (0.0000)	0.7500 (0.0000)	0.7792 (0.0000)					
Hierarchical	Latin	0.5618 (0.0000)	1.0000 (0.0000)	0.7194 (0.0000)	0.5667 (0.0000)	0.4395 (0.0000)	0.0243 (0.0000)	0.0056 (0.0000)	2
	modern Italian	0.4382 (0.0000)	0.9750 (0.0000)	0.6047 (0.0000)					
SOM	Latin	0.8116 (0.0010)	0.8616 (0.0055)	0.8358 (0.0031)	0.8120 (0.0030)	0.6195 (0.0069)	0.2987 (0.0070)	0.3825 (0.0078)	2
	modern Italian	0.8126 (0.0061)	0.7500 (0.0000)	0.7800 (0.0028)					
K-Medians	Latin	0.8113 (0.0000)	0.8600 (0.0000)	0.8350 (0.0000)	0.8111 (0.0000)	0.6215 (0.0000)	0.2967 (0.0000)	0.3803 (0.0000)	2
	modern Italian	0.8108 (0.0000)	0.7500 (0.0000)	0.7792 (0.0000)					

We observe that our framework, which is the combination of run-length features and GA-ICDA$^+$ clustering method, performs successfully, overcoming all the other clustering methods (see Table 1). In fact, GA-ICDA$^+$ obtains the perfect distinction between Latin and modern Italian documents, with a number of clusters equal to 2, precision, recall and f-measure values of 1.00 for both Latin and modern Italian language classes, purity, NMI and ARI values of 1.00 and an entropy value of 0.00. Furthermore, standard deviation values are always zero, demonstrating the stability of the result. It is interesting to observe as GA-IC algorithm is not able to well discriminate the languages. Although the number of found clusters is exactly 2, the f-measure values are 0.83 for Latin and 0.78 for modern Italian, the purity value is 0.81, the NMI value is quite low and equal to 0.30, together with the value of ARI which is 0.38 and the high value of entropy which is 0.62. This means that the found clusters contain mixed Latin and modern Italian documents. The GA-ICDA procedure performs considerably better than GA-IC for this task. In fact, it exhibits f-measure values of 0.95 for Latin and 0.94 for modern Italian, a purity value of 0.94, a entropy value of 0.22 and NMI and ARI values of respectively 0.74 and 0.79. It indicates that GA-ICDA is more apt to deal with document data than GA-IC. However, the best result is given from GA-ICDA$^+$, demonstrating the efficacy of the performed modifications. About the other algorithms, we can observe that a pure bottom-up strategy like hierarchical clustering is not able to outperform the GA-IC, GA-ICDA and GA-ICDA$^+$ evolutionary strategies. In fact, it reaches f-measure values of 0.72 and 0.60 for respectively Latin and modern Italian, a purity value of 0.57, a entropy value of 0.44 and very low NMI and ARI values of respectively 0.02 and 0.006. It is also worth to note that the results of GA-ICDA, adopting together an evolutionary method and a bottom-up refinement procedure, are better than both the pure evolutionary procedure of GA-IC and the pure bottom-up strategy of hierarchical clustering. It demonstrates the efficacy of the combination of both the evolutionary and bottom-up methods in document clustering. The SOM results are very similar to the results obtained from GA-IC. In fact, the f-measure values are equal to 0.83 for Latin and 0.78 for modern Italian, the purity and entropy values are respectively 0.81 and 0.62, the NMI and ARI values are quite low and respectively 0.30 and 0.38. K-Medians also obtains results which are similar to the results of SOM and GA-IC, with a f-measure value of 0.83 for Latin and 0.78 for modern Italian, purity, NMI and ARI values of respectively 0.81, 0.30 and 0.38 and a very high entropy value of 0.62. It indicates that GA-IC, SOM and K-Medians are trapped into a recurrent solution consisting of mixed clusters of documents in Latin and modern Italian languages.

5 Conclusions

The paper introduced a new framework for the discrimination between documents written in Latin and modern Italian languages. It is characterized by the position of each script letter in the baseline, derived by its energy profile, for mapping into uniformly coded text. The statistical analysis of the coded text, represented as an image, is performed by the run-length matrix calculation for texture feature extraction. The obtained feature vectors revealed satisfactory dissimilarity of the documents in different languages. Such a dissimilarity is the basis for successfully document clustering by the extension of a state-of-the-art classification tool GA-ICDA$^+$. Experimental results demonstrated the superiority of the new framework with respect to the other clustering methods. Future work will extend the experiment to larger databases and multiple types of language feature representations.

Acknowledgments. This work was partially supported by the Grant of the Ministry of Education, Science and Technological Development of the Republic Serbia, as a part of the project TR33037. Authors are fully grateful to Ms. Zagorka Brodić, professor of French and Serbo-Croatian languages, for the helpful discussions about Latin and Italian languages.

References

1. Amelio, A. and Pizzuti, C.: A new evolutionary-based clustering framework for image databases. In: Image and Sign. Proc., June 30-July 2, Cherbourg, Normandy, France, 8509:322-331 LNCS, Springer, 2014.

2. Andrews, N. O. and Fox, E. A.: Recent Developments in Document Clustering. Technical report, Computer Science, Virginia Tech.
3. Brodić, D., Amelio, A., Milivojević, Z. N.: Characterization and Distinction Between Closely Related South Slavic Languages on the Example of Serbian and Croatian. In: Comp. Anal. of Images and Patterns, 2-4 September, Valletta, Malta, 9256:654-666 LNCS, Part I, Springer, 2015.
4. Brodić, D., Milivojević, Z.N., Maluckov, C.A.: Recognition of the script in serbian documents using frequency occurrence and co-occurrence analysis. The Scientific World Journal, 896328:1-14, 2013.
5. Calabrese, A.: On the Evolution of the short high vowel of Latin into Romance, in A. Perez-Leroux & Y Roberge (eds.) Romance Linguistics. Theory and Acquisition. Amsterdam, John Benjamins, 63-94, 2003.
6. Chu, A., Sehgal, C.M., Greenleaf, J.F.: Use of gray value distribution of run lengths for texture analysis. Pattern Recognition Letters 11(6):415-419, 1990.
7. Dasarathy, B.R., Holder, E.B.: Image characterizations based on joint gray-level run-length distributions. Pattern Recognition Letters, 12(8):497-502, 1991.
8. Galloway, M.M.: Texture analysis using gray level run lengths. Computer, Graphics and Image Processing 4(2):172-179, 1975.
9. Joshi, G.D., Garg, S., Sivaswamy, J.: A generalised framework for script identification. *IJDAR*, 10(2):55-68, 2007.
10. https://sites.google.com/site/documentanalysis2015/latin-italian-database.
11. Manning, C.D., Raghavan P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, 2008.
12. Powers, D. M. W.: Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies 2(1):37-63, 2011.
13. Saarikoski, J., Laurikkala, J., Järvelin, K., Juhola, M.: Self-Organising Maps in Document Classification: A Comparison with Six Machine Learning Methods. In: 10th Int. Conf., ICANNGA, 14-16 April, Ljubljana, Slovenia, Part I 6593:260-269 LNCS, Springer, 2011.
14. Santos, J. M. and Embrechts, M.: On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In: 19th International Conference on Artificial Neural Networks: Part II, 14-17 September, Limassol, Cyprus, Springer-Verlag, Berlin, Heidelberg, 175-184.
15. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining, 20-23 August, Boston, MA, USA, 2000.
16. Tan, X.: Texture information in run-length matrices. IEEE Trans. Image Proc. 7(11):1602-1609, 1998.
17. De Vries, C.M., Geva, S. and Trotman, A.: Document clustering evaluation: Divergence from a random baseline. CoRR, abs/1208.5654, 2012.
18. Hu, X. and Yoo, I.: A comprehensive comparison study of document clustering for a biomedical digital library medline. In: 6th ACM/IEEE-CS Joint Conference on, 11-15 June, Chapel Hill, NC, USA, 220-229, 2006.
19. Zhao, Y., Karypis, G., Fayyad, U.: Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery, 10(2):141-168, 2005.
20. Zramdini, A., Ingold, R.: Optical font recognition using typographical features. IEEE Trans. Pattern Analysis and Machine Intelligence, 20(8):877-882, 1998.

The First Cross-Script Code-Mixed Question Answering Corpus

Somnath Banerjee[1], Sudip Kumar Naskar[1], Paolo Rosso[2], and Sivaji Bandyopadhyay[1]

[1] Computer Science and Engineering Department, Jadavpur University, India
sb.cse.ju@gmail.com,{sudip.naskar,sbandyopadhyay}@cse.jdvu.ac.in
[2] PRHLT Reearch Center, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract. In this paper, we formally introduce the problem of cross-script code-mixed question answering (QA) and we elaborate the corpus acquisition process and an evaluation strategy related to the said problem. Today social media platforms are flooded by millions of posts everyday on various topics. This paper emphasizes the use of such ever growing user generated content to serve as information collection source for the QA task on a low-resource language for the first time. A majority of these posts are multilingual in nature and many of them involve code mixing. The multilingual aspect of social media content is reflected in the use of multilingual words as well as in the writing script. For the ease of use multilingual users often pose questions in non-native script. Focusing on this current multilingual scenario, code-mixed cross-script (i.e., non-native script) data give rise to a new problem and present serious challenges to automatic QA. In the work presented in this paper, Bengali is considered as the native language while English is considered to be the non-native language. However, the dataset construction approach presented in this paper is generic in nature and could be used for any other language pair. Apart from introducing this novel problem, this paper highlights corpus development process and a suitable evaluation framework.

Keywords: Question Answering, Code Mixing, Code Switching, Cross-script, social media

1 Introduction and Related Work

Code-mixing refers to the phenomenon where lexical items and grammatical features from two languages appear in one sentence. The use of code-mixing is spreading widely in informal text communications such as newsgroups, tweets, blogs, and other social media platforms. Sometimes it is used to refer to relatively stable informal mixtures of two languages, such as Spanglish, Franponais or Portuñol. Nowadays in social media people tend to share everything under the sun. Social media users often share their travel experiences as well as seek travel suggestions from their social networks. Similarly sports events are among the mostly discussed topics in social media. People post live updates of ongoing sports events such as Football World Cup, Champions League, T20 Series, etc. This results in potentially rich resources for languages which are less computerized.

In bilingual or multilingual countries like India, speakers often incorporate lexical items, phrases, and clauses from more than one language into their spoken or written communication act. This results in words or phrases from different languages in the same sentence or utterance. This phenomenon is referred to as code-mixing. Although this phenomenon has been studied extensively in formal and spoken context, the research community in natural language processing (NLP) has just started paying sincere attention to code-mixing due to its prevalence of use in electronic communication mainly in the social media. English is predominantly the most used language on the internet; Indians also use English extensively while surfing the internet. Even they (phonetically) use the Roman script instead of using their own native scripts. Another important reason behind the use of the English language and the Roman script may be the keyboards which are in the non-native Roman script, and Indian internet users are more comfortable using that keyboard rather than the on-screen native script keyboard or a combination of keys which generate native alphabets. Every natural language is generally written using a particular script which is referred to as the native script for that language. All other scripts which are not used in writing the language can be referred to as the non-native script with respect to that language. For example, the English language is written in the Roman script. Thus, Roman script is the native script for English, however Bengali script is a non-native script for English. We refer to the phenomenon of using a non-native script phonetically for writing native words as cross-script. For example, if a Bengali user writes Bengali words in Bengali script, that is considered as using native script. However, if he writes Bengali words in Roman script or English words in Bengali script, then he is making use of cross-script.

Being a classic application of NLP, QA has practical applications in various domains such as education, health care, personal assistance, etc. Presently, QA is a well addressed research problem and several QA systems are available with reasonable accuracy. A number of QA systems were developed for European languages particularly for English ([1], [2],[3],[4]), Middle Eastern languages ([5],[6],[7]) and Asian languages, e.g., Japanese ([8],[9]) Chinese ([10],[11]). In this paper, we introduce a new research problem in the context of QA research cross-script code-mixed QA.

The rest of the paper is organized as follows. Section 2 states the code-mixed cross-script QA problem. We discuss corpus acquisition in Section 3. The proposed corpus annotation process and corpus statistics are described in Section 4 and Section-5, respectively. We present the evaluation scheme in Section 6. Section 7 concludes the paper.

2 Problem Statement

Problem Statement: *Building a question answering system which takes cross-script (non-native) code-mixed questions as information request, processes a cross-script code-mixed text corpus and provides an (or a list of) exact answer(s) as information response.*

We introduce this novel research problem for the following reasons:

1. Multilingual non-native English speakers predominantly use the Roman script in social media platforms during their conversations even while the written communication takes place entirely in a native language (i.e., not English).
2. To make the written communication more fascinating, borrowing foreign words from different languages is very common in social media communication and this is a growing trend.
3. The ever increasing posts in many less-computerized languages could serve as a potential source of digital content for language research.
4. The research community need to move towards the next generation search engine that boosts the necessity of developing QA system for less-resourced languages.

This paper presents a cross-script code-mixed QA corpus for Bengali; however, this context is very common with other non-English languages, e.g. Spanish, French, etc. Despite the advances in QA research and the fact that Bengali is one of the most spoken languages, very little work ([12],[13],[14]) has been conducted in QA for Bengali so far. Language identification in the code-mixing scenario has been addressed extensively in shared tasks in EMNLP-2014[3] and FIRE-2014[4] and in few other research works [15],[16],[17],[18]. However, to the best of our knowledge, no work has been conducted so far on the novel problem addressed in this paper.

3 Corpus Acquisition

Because of the following characteristics of social media, we consider social media content for code-mixing cross-script QA corpus:

i) Substantial and ever increasing user base.

ii) A sizable volume of informal text data are added on various domains on a daily basis.

iii) Various APIs are available to access social media data.

iv) Most likely source of getting code-mixed data.

Even though acquiring a sizable volume of the code-mixed cross-script data is not a tough task, our work on developing a QA system for code-mixed cross-script data is at its initial stages. Therefore, we have collected a small set of data which could be increased in future following with a similar approach. Research

[3] http://emnlp2014.org/workshops/CodeSwitch/call.html

[4] http://fire.irsi.res.in/fire/home

in QA system primarily requires three data resources: (i) question which is asked to get a piece of information, (ii) answer to an asked question as a response, and (iii) potential sources of the answers from which a QA system can directly or indirectly infer an answer to a question. We describe the acquisition of these resources in this section. For the present study, we restricted our focus to the tourism and the sports domains which are among the most popular domains in the social media. Social media data on other domains could be acquired with a similar approach presented here. In the code-mixed cross-script QA scenario, the resource development involves two separate processes: (i) collecting social media text for the desired domains; and (ii) question acquisition and answer annotation.

3.1 Message(text) Acquisition

For the document collection we consider the social media as it is the most likely potential source of code-mixed cross-script data. We acquired all the messages from different social media platforms, e.g., twitter, blogs, forums, etc. For the sports domain, we selected social media posts on recently held 10 exciting cricket matches. Ten popular tourist spots in India were selected for tourism domain. Tweepy API and an in-house focused crawler were employed for collecting tweets, blogs, and forum posts. For collecting only code-mixed data, we set a language mix ratio (i.e., non-native:native) which is computed by employing a language identifier whose accuracy, as reported in [19], is 92.4%. Language mixing ratio (LMR) is employed for collecting only code-mixed data. The language mixing ratio has been set to 0.2 after manually verifying a small set of crawled data. Therefore, a message post is included in the corpus when at least 16.67% (i.e. 1 in 6) of the words belong to the non-native language.

Examples of valid Message:

a) Message: SA\O ja\B run\E koreche\B aj\B BD\O parbe\B ki\B ?\O

LMR $= \frac{\#non-native}{\#native} = \frac{\#English-words}{\#Bengali-words} = \frac{1}{5} = 0.2(>= 0.2)$

b) Message: Mashrafe\O well\E try\E but\E ki\B r\B kora\B jabe\B ...\O captain\E !!!\O

LMR $= \frac{\#non-native}{\#native} = \frac{\#English-words}{\#Bengali-words} = \frac{4}{4} = 1(>= 0.2)$

The language identifier, as reported in [19], does not identify named entities. Considering the fact that the answer to a factoid question is always a named entity, we filtered out the messages under human supervision which do not contain any named entity. Thus, we finalized 299 posts as messages out of the 334 messages which were initially selected by the language identifier and the LMR ratio.

3.2 Question Acquisition

The question preparation task is more challenging than the message acquisition and requires more human involvement. Our prime target was to involve as many question setters as possible to reduce bias. A cloud-based service was built and requests were sent to the undergraduate students of the university. Two groups, namely sports-domain group (SG) and tourism-domain group (TG) with 15 students each were formed from thirty students who agreed for the question annotation task. Ten topics on sports domain were provided to each member of SG and they were asked to submit at least 10 questions on each topic. The submitted questions were stored in the web server along with the messages associated with the topic. After receiving these questions, we kept only the questions having code-mixed nature and satisfying the LMR criterion. Subsequently, the annotators were asked to find out the answer to their legitimate questions from the stored messages. An analogous procedure was followed for TG also.

4 Annotation

For document management and storing, EXtensible Markup Language (XML) was chosen because of its popularity and ease of understanding. The QA annotation framework which was adopted in this work is depicted in Fig. 1. The tagset defined in Table 1 was used for three purposes: document information, message annotation and QA annotation. We will format the corpus in Text Encoding Initiative[5](TEI) in future.

Table 1. Corpus tagset

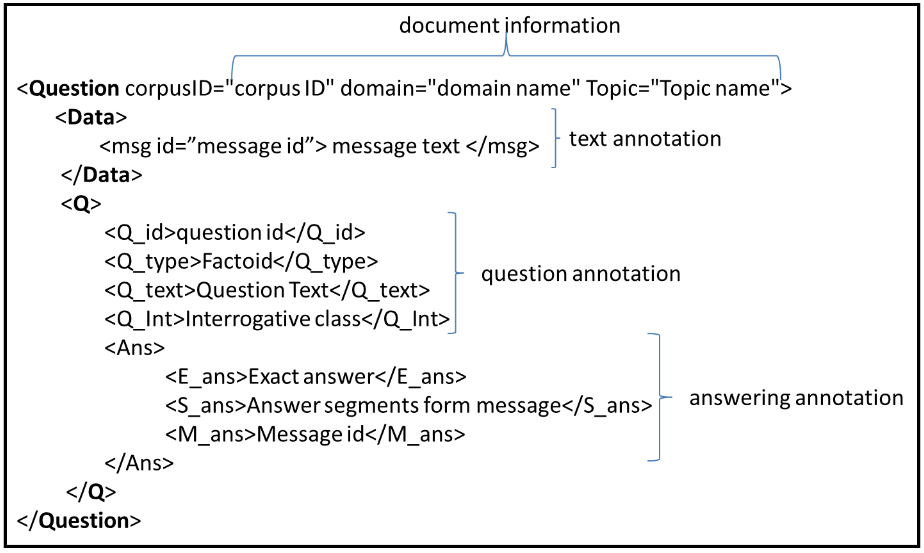
Tag	Definition	Tag	Definition
Question	Document body	CorpusID	Corpus id number
Domain	Domain name	Topic	Topic name
Data	Data section	Q	Question
Q_id	Question unique number	Q_type	Question type, e.g., Factoid, Procedural
Q_text	Code-mixed NL question	Q_Int	Interrogative class
Ans	Answer	E_ans	Exact answer
S_ans	Segment answer	M_ans	Message Id of a message that contains answer
Msg	Public posts as messages		

A document in the corpus comprises of data section and question section. The data section contains the public posts collected from social media. Each public post is referred to as a message and described in the $< msg >$ tag. Each message is assigned a unique number, i.e., msg_Id. The factoid questions follow the data section. Each question is marked by the Q tag, (i.e., $< Q >$ and $< /Q >$). Like each message, every question is also assigned a unique question identifier. The question type (*Q_type*) denotes the type of a question such as factoid, procedural, etc. The code-mixed cross script question is enclosed by the *q_text* tag.

Interrogative types of questions are very much useful for answer extraction and validation. On the basis of syntactic structure, Bengali interrogatives are

[5] http://www.tei-c.org/index.xml

Fig. 1. Document Template



```
<Question corpusID="corpus ID" domain="domain name" Topic="Topic name">
     <Data>
          <msg id=”message id”> message text </msg>
     </Data>
     <Q>
          <Q_id>question id</Q_id>
          <Q_type>Factoid</Q_type>
          <Q_text>Question Text</Q_text>
          <Q_Int>Interrogative class</Q_Int>
          <Ans>
               <E_ans>Exact answer</E_ans>
               <S_ans>Answer segments form message</S_ans>
               <M_ans>Message id</M_ans>
          </Ans>
     </Q>
</Question>
```

classified into three categories - single interrogative (SI), dual interrogative (DI) and compound interrogative (CI) [12]. The interrogative type (i.e., SI, DI, and CI) of a question gives a clue about the number information of the candidate answer.

The answer to a question is annotated by the Ans tag. The exact answer is given in *E_ans* tag. The Segment answer (*S_ans*) tag refers to the portion or segment of the message text which provides the answer. The message id from which the exact answer can be found is given in the message answer (*M_ans*) tag. The segment answer tag and message tag could be thought of as supporting information for the exact answer.

5 Corpus Statistics

The statistics of the messages, i.e., public posts and questions in the corpus for the two different domains, namely Sports and Tourism, are given in Table 2. Altogether 299 code-mixed cross-script messages were collected of which 183 and 116 messages are from the tourism and sports domains respectively. 506 code-mixed cross-script questions were acquired of which 314 questions are from the tourism domain and 192 questions belong to the sports domain. Average number of messages per document (Avg. M/D in Table 2)is higher for the tourism domain than for the sports domain. Average number of questions generated per document (Avg. Q/D in Table 2) is higher for the tourism domain than for the sports domain accordingly.

Table 2. Corpus statistics

Domain	Documents(D)	Messages(M)	Questions(Q)	Avg. M/D	Avg. Q/D
Tourism	10	183	314	18.3	31.4
Sports	10	116	192	19.2	19.2
Overall	20	299	506	14.95	25.3

6 Proposed Evaluation

Along with the corpus development, we also propose an evaluation scheme to evaluate the code-mixed QA performance which is suitable to our corpus annotation. In the annotated corpus an answer is basically structured as [*Answer String* (AS), *Message Segment* (MS), *Message ID* (MId)] triplet, where-

– AS is the one of the exact answers (EA) and must be an NE in this case,

– MS is the supported text segment for the extracted answer, and

– MId is the unique identifier of the message that justifies the answer.

The evaluation methodology was designed taking into consideration the following issues:

i) The QA system has the provision of not answering, i.e., no answer option (NAO).

ii) The answer returned should be the exact answer to the question.

iii) The exact answer must be a Named Entity.

iv) The system has to return a single exact answer. In case there exists more than one correct answer to a question, the system needs to provide only one of the correct answers.

While designing the evaluation strategy, our primary focus was on "responsiveness" and "usefulness" of each answer. Each answer has to be manually judged by native speaking assessors. Each answer [AS, MS, MId] triplet is assigned a score in a five-valued (range 0.0-1.0) scale which is weighted correctness measure using hard-coded weights and marked with exactly one of the following judgments depicted in Table 3:

- **Incorrect:** The AS does not contain EA (i.e., responsive but not useful)
- **Unsupported:** The AS contains correct EA, but MS and MId do not support the EA (i.e., missing usefulness)
- **Partial-supported:** The AS contains the correct EA with correct MId, but MS does not support EA
- **Correct:** The AS provides the correct EA with correctly supporting MS and MId (i.e., "responsive" as well as "useful").
- **Inexact:** The supporting MS and MId are correct, but the AS is wrong.

The QA evaluation forums such as TREC[6], CLEF[7], etc. proposed accuracy, c@1[20], and Mean reciprocal rank (MRR) [21] as evaluation metrics for the

[6] http://trec.nist.gov/

[7] http://www.clef-initiative.eu/

Table 3. Judgment Scale

Judgment	**AS**	**MS**	**MId**	**Score**
Incorrect (W)	X	X	X	0.00
Inexact(I)	X	✓	✓	0.25
Unsupported (U)	✓	X	X	0.50
Partial-supported (P)	✓	X	✓	0.75
Correct (C)	✓	✓	✓	1.00

monolingual and cross-lingual QA. In order to maintain the consistency with the state-of-the-art QA evaluation metrics, we also suggest the use of accuracy and c@1 for the code-mixed cross-script QA task. As the prepared corpus contains only one correct answer (as opposed to a list of exact answers) for every question, MRR is not useful for evaluation on the said dataset. Just as in the past ResPubliQA[8] campaigns, systems have the option of withholding the answer to a question because they are not sufficiently confident that it is correct (i.e., NAO). As per ResPubliQA, the inclusion of NAO improves the system performance by reducing the number of incorrect answers.

Now, C@1 = $\frac{1}{N}(N_r + N_u . \frac{N_r}{N_u})$

Accuracy = $\frac{N_r}{N}$

C@1 = Accuracy; if $N_u = 0$

Where, N_r = number of right answers.

N_u = number of unanswered questions

N = total questions

Correct, Partially-supported and Unsupported answers provide the exact answers only.

Therefore, $N_r = (\#C + \#U + \#P)$

Considering the importance of supporting segment, we introduce a new metric "answer-support performance" (ASP) which measures the answer correctness and which is defined as follows:

$ASP = \frac{1}{N}(c \times 1.0 + p \times 0.75 + i \times 0.25)$

where, c, p and i denote total number of correct, partially-supported and inexact answers respectively.

7 Conclusions

In this paper we presented a novel research problem - cross-script code-mixed QA. Our major contributions include (i) proposing an annotation scheme, ii) creating a dataset which is the first resource of its kind, and (iii) proposing an evaluation strategy that is suitable to our corpus annotation. Bearing in mind the small dataset, the proposed evaluation methodology and created dataset will be helpful for the QA research and development community, particularly those who want to address code-mixed cross-script QA.

[8] http://nlp.uned.es/clef-qa/repository/resPubliQA.php

Acknowledgements

We acknowledge the support of the Department of Electronics and Information Technology (DeitY), Government of India, through the project "CLIA System Phase II". The work of the third author was in the framework of the SomEMBED MINECO TIN2015-71147-C2-1-P research project.

References

1. Buscaldi, D., Rosso, P., Gómez, J.M., Sanchis, E.: Answering Questions with an n-gram based Passage Retrieval Engine. In: Journal of Intelligent Information Systems, 34:113-134 (2010)
2. Brill, E., Dumais, S., Banko, M.:. An analysis of the AskMSR question-answering system. In: Empirical methods in natural language processing-Volume 10, pp. 257-264, Association for Computational Linguistics (2002)
3. Zheng, Z.: AnswerBus question answering system. In: International conference on Human Language Technology Research, pp. 399-404, Morgan Kaufmann Publishers Inc. (2002)
4. Ittycheriah, A., Franz, M., Zhu, W. J., Ratnaparkhi, A., Mammone, R. J.: IBM's Statistical Question Answering System. In: TREC (2000)
5. Mohammed, F. A., Nasser, K., Harb, H. M.: A knowledge based Arabic question answering system (AQAS). In: ACM SIGART Bulletin, 4(4), 21-30 (1993)
6. Kanaan, G., Hammouri, A., Al-Shalabi, R., Swalha, M.: A new question answering system for the Arabic language. In: American Journal of Applied Sciences, 6(4), 797 (2009)
7. Hammo, B., Abu-Salem, H., Lytinen, S.: QARAB: A question answering system to support the Arabic language. In: ACL-02 workshop on Computational approaches to semitic languages, pp. 1-11, Association for Computational Linguistics (2002)
8. Sakai, T., Saito, Y., Ichimura, Y., Koyama, M., Kokubu, T., Manabe, T.: ASKMi: A Japanese Question Answering System based on Semantic Role Analysis. In: RIAO, pp. 215-231 (2004)
9. Isozaki, H., Sudoh, K., Tsukada, H.: NTTs japanese-english cross-language question answering system. In: NTCIR Workshop 5 Meeting, pp. 186-193 (2005)
10. Yongkui, Z. H. A. N. G., Zheqian, Z. H. A. O., Lijun, B. A. I., Xinqing, C. H. E. N.: Internet-based Chinese Question-Answering System. In: Computer Engineering, 15 (2003)
11. Sun, A., Jiang, M., He, Y., Chen, L., Yuan, B.: Chinese question answering based on syntax analysis and answer classification. In: Acta Electronica Sinica, 36(5) (2008)
12. Banerjee, S., Bandyopadhyay, S.: Bengali Question Classification: Towards Developing QA System. In: SANLP-COLING, IIT,Mumbai,India (2012)
13. Banerjee, S., Lohar, P., Naskar, S. K., Bandyopadhyay, S.: The First Resource for Bengali Question Answering Research. In: PolTAL-2014. Poland. In Advances in Natural Language Processing, pp. 290-297. Springer International Publishing (2014)
14. Banerjee, S., Naskar, S. K., Bandyopadhyay, S.: BFQA: A Bengali Factoid Question Answering System. In: Text, Speech and Dialogue (TSD), pp. 217-224. Springer International Publishing, Czech Republic (2014)
15. Gupta, P., Bali, K., Banchs, R., Choudhury, M., Rosso, P.: Query Expansion for Mixed-script Information Retrieval. In: The 37th Annual ACM SIGIR Conference, SIGIR-2014, Gold Coast, Australia, June 6-11, pp. 677-686 (2014)

16. King, B., Abney, S.: Labeling the languages of words in mixed-language documents using weakly supervised methods. In: NAACL-HLT, pages 1110–1119 (2013)
17. Barman, U., Wagner, J., Chrupala, G., Foster, J.: Identification of languages and encodings in a multilingual document. In: EMNLP (2014)
18. Choudhury, M., Chittaranjan, G., Gupta, P., Das, A.: Overview of FIRE 2014 Track on Transliterated Search. In: FIRE (2014)
19. Banerjee, S., Kuila, A., Roy, A., Naskar, S. K., Bandyopadhyay, S., Rosso, P.: A Hybrid Approach for Transliterated Word-Level Language Identification: CRF with Post Processing Heuristics. In: Forum for Information Retrieval Evaluation, pp. 54-59, ACM Digital Publication (2014)
20. Peñas, A., Rodrigo, A.: A Simple Measure to Assess Non-response. In: 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT 2011). Portland, Oregon, USA (2011)
21. Voorhees, E.M.: The TREC-8 question answering track report. In: 8th Text Retrieval Conference (TREC), Gaithersburg, Maryland, USA, pp. 77-82 (1999)