Journal of Artificial Intelligence Research 55 (2016) 1-15 Submitted 11/15; published 01/16

Introduction to the Special Issue on Cross-Language Algorithms and Applications

Marta R. Costa-jussà MARTA.RUIZ@UPC.EDU
TALP Research Center
Universitat Politècnica de Catalunya
Jordi Girona 1–3, 08034, Barcelona

Srinivas Bangalore SBANGALORE@INTERACTIONS.NET
Interactions Labs
41 Spring Street,
Murray Hill, NJ 07974, USA

Patrik Lambert PATRIK.LAMBERT@UPF.EDU
Computational Linguistics Group
Universitat Pompeu Fabra
Roc Boronat 138, 08018 Barcelona, Spain

Lluís Màrquez LMARQUEZ@QF.ORG.QA
Qatar Computing Research Institute
Hamad Bin Khalifa University
Tornado Tower (10th floor), PO.Box 5825,
West Bay, Doha, Qatar

Elena Montiel-Ponsoda ELENA.MONTIEL@UPM.ES
Ontology Engineering Group
Universidad Politécnica de Madrid
Campus de Montegancedo s/n, Boadilla del Monte,
28660 Madrid

Abstract

With the increasingly global nature of our everyday interactions, the need for multilingual technologies to support efficient and effective information access and communication cannot be overemphasized. Computational modeling of language has been the focus of Natural Language Processing, a subdiscipline of Artificial Intelligence. One of the current challenges for this discipline is to design methodologies and algorithms that are cross-language in order to create multilingual technologies rapidly. The goal of this JAIR special issue on Cross-Language Algorithms and Applications (CLAA) is to present leading research in this area, with emphasis on developing unifying themes that could lead to the development of the science of multi- and cross-lingualism. In this introduction, we provide the reader with the motivation for this special issue and summarize the contributions of the papers that have been included. The selected papers cover a broad range of cross-lingual technologies including machine translation, domain and language adaptation for sentiment analysis, cross-language lexical resources, dependency parsing, information retrieval and knowledge representation. We anticipate that this special issue will serve as an invaluable resource for researchers interested in topics of cross-lingual natural language processing.

1. Introduction

Due to the increasingly global nature of our society, it is commonplace for each of us to encounter information in a multitude of languages and to communicate across languages in our everyday lives. The rapid growth of multilinguality in our information-driven society is reflected in the number of languages being used on the Internet. In about a decade, the Internet has transformed from being a predominantly English information source to the linguistically variegated information source that it is today. Multilingual access and processing pose novel challenges to the core Artificial Intelligence discipline of speech and natural language processing which, when solved, can provide transformational technologies for information access to a broader population.

The complexity of processing multiple languages in a computational model emerges not only due to different syntactic structures for the same concept, but also from different underlying conceptual structures. These challenges necessitate the development of cross-language natural language processing tools that are able to translate or link these structures and concepts across different languages.

Cross-language extensions of popular applications and tasks such as information retrieval, question answering, sentiment analysis and lexical disambiguation, among others, have been developed to respond to the present needs of the global society. Similarly, multilingual resources and novel ways to represent multilingual knowledge have emerged and spread in recent years. Researchers and developers have made considerable advances in these applications by leveraging relevant data available in diverse languages. With multilingual processing becoming a key research issue, and given its interdisciplinary nature, a range of approaches from both linguistic and statistical perspectives have been explored in order to create viable cross-lingual technology.

Interestingly, the challenges of cross-lingual natural language processing appeals to both academic and industrial research. It provides new business opportunities by breaking down language barriers that fragment the potential market. These opportunities include, for instance, the possibility for companies and institutions to learn about what users at different locations think of their products (cross-language sentiment analysis), to perform document search in multiple languages (cross-language information retrieval and question answering), to link knowledge bases of clients using different languages (cross-language knowledge representation), or to expand the market to several linguistically diverse markets (machine translation and localization).

The importance of cross-lingual natural language processing can be clearly seen from the attention it has received at recent workshops, invited talks, and publications at the community's premier conferences—Association of Computational Linguistics (ACL), European Association of Computational Linguistics (EACL), North American Association of Computational Linguistics (NAACL), Empirical Methods in Natural Language Processing (EMNLP), International Conference on Computational Linguistics (COLING), Extended Semantic Web Conference (ESWC), International Semantic Web Conference (ISWC), International Conference on Language Resources and Evaluation (LREC), and International Conference on Knowledge Capture (KCAP). In the past eighteen months, about one in every five papers at these conferences was related to cross-language algorithms and applications (figures vary from 17% to 25%, with an average of 20.4%). A recently published book has focused on multilingual natural language processing techniques (Bikel & Zitouni, 2012) and further highlights the importance of this research area. This book comprehensively presents, in more than 600 pages, the basic theory and the most relevant techniques for 16 different aspects of multilingual natural language processing.

1.1 Cross-Language Algorithms and Applications Special Issue

This special issue is intended to provide a broad view of some of the recent advances and current research directions being pursued in the area of multilingual natural language processing from the linguistic, computational and language resource creation perspectives. Active research from multiple recent workshops in the area (e.g., HyTra, see Costa-juss, Banchs, Rapp, Lambert, Eberle & Babych, 2013; WMT, see Bojar, Chatterjee, Federmann, Haddow, Huck, Hokamp, Koehn, Loncheva, Monz, Negri, Post, Scarton, Specia & Turchi, 2015; CLEF, see Forner, Moller, Paredes, Rosso & Stein, 2013; Promise, see Bener, Minku & Turhan, 2014; MSW, see Gracia, MacGrae & Vulcu, 2015; and AKBC, see Suchanek, Riedel, Singh & Talukdar, 2012) spanning cross-language natural language applications, tools of crowdsourcing for resource creation, and deeper relationships between bridging language barriers and other modalities of perception are summarized in this special issue.

In response to our solicitation of papers in late 2014, we received 34 research papers on a variety of cross-lingual technologies applied to language processing tasks. Each of these papers was carefully reviewed by at least three reviewers drawn from a pool of over 100 reviewers. Based on their reviews and extensive follow up discussions, we selected 8 high quality papers that offer exciting research directions which span a range of topics in cross-lingual language processing including machine translation, domain and language adaptation for sentiment and cross-language lexical resources, dependency parsing, information retrieval and knowledge representation. This introduction covers exactly the eight papers that were accepted for the special issue during the official timeline. The papers are summarized in Section 2, organized by topic. In an attempt to accommodate a broader sample of interesting papers on cross-language algorithms and applications, other papers appropriate for the topic that did not meet the special issue deadlines will be also added to the same JAIR web page[1] when accepted by the journal.

Finally, we intended the special issue to not be a disconnected melange of success stories, but provide an underlying theme that unifies the research directions in this area. We hope that this collection provides the reader an opportunity to observe similarities and differences across topics, algorithms and applications. We anticipate that the scientific community will view cross-lingual speech and language processing as a fertile and productive field of research, that has the potential for developing science and technologies which will have a lasting impact on our everyday lives.

2. Special Issue Overview

In this section, we survey the topics that cover the papers in this special issue and the papers themselves. These topics are machine translation, domain and language adaptation

1. http://www.jair.org/specialtrack-claa.html

for sentiment analysis and cross-language lexical resources, dependency parsing, information retrieval and knowledge representation.

2.1 Machine Translation

A key technology for the multilingual information society is *Machine Translation* (MT)—where a source language speech or text is automatically converted into a target language speech or text.

With Web content being generated in multiple languages and with Internet users becoming linguistically diverse, machine translation provides the cheapest and quickest way to understand this multilingual information. Besides being a core application necessary for a multilingual world, machine translation has been shown to be a highly relevant *component* technology in many cross-language natural language processing tasks, such as sentiment analysis, information retrieval and knowledge representation.

Historically, machine translation approaches can be categorized into rule-driven and data-driven approaches. Rule-based machine translation (Hutchins & Sommers, 1992) requires deep linguistic knowledge of the language pairs involved in the translation and a significant amount of human labor. Since the rules are based on linguistic intuitions, it is easier to identify issues and extend them. More recently, there have been data-driven approaches that learn translation models with minimal human supervision from very large bilingual parallel corpora—texts where source text is paired with the target text (Sánchez-Martínez & Forcada, 2009). Data-driven translation systems find the most probable target text given the source text. These systems have been extended to phrase-based, syntax-based, hierarchical phrase-based and neural-based systems in order to capture longer contexts in a sentence. Phrase-based systems (Koehn, Och, & Marcu, 2003) use sequences of words as bilingual units, called phrases, whose probability is computed through a log-linear combination of feature functions including the translation and language models. Syntax-based systems use syntactic units extracted by parse trees (Quirk, Menezes, & Cherry, 2005). Hierarchical phrase-based systems combine phrase-based and syntax-based approaches by using synchronous context-free grammars (Chiang, 2007). Neural-based end-to-end translation systems typically use a *encoder-decoder* approach to learn embedded representations of the input sentence (encoder), which are then used as context to generate the words in the translation (decoder) (Bahdanau, Cho, & Bengio, 2015).

The boundaries between rule-based and statistical machine translation have narrowed through the proposals of hybrid machine translation systems (Costa-jussà, 2015). In this special issue, there are two research papers in machine translation that combine rules, statistics and machine learning.

Integrating Rules and Dictionaries from Shallow-Transfer Machine Translation into Phrase-Based Statistical Machine Translation by *Sánchez-Cartagena, Pérez-Ortiz and Sánchez-Martínez* presents a hybrid approach to machine translation by integrating rules and dictionaries from a shallow-transfer system (in particular, Apertium, see Armentano-Oller & Forcada, 2006) into a phrase-based system (in particular, Moses, see Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin & Herbst, 2007). Deep linguistic knowledge from rule-based systems is transferred into the statistical system. This integration is especially useful when the parallel corpus available for training the phrase-based system is small or when translating out-of-domain texts. The authors discuss different methods for enriching translation tables (including previous work by the same authors and other references). Finally, it is worth mentioning that the approach that used Apertium's rules was one of the best-ranked systems in the WMT 2011 international evaluation campaign for Spanish-English (Sánchez-Cartagena, Sánchez-Martínez, & Pérez-Ortiz, 2011).

Two relevant features of this paper are: (*i*) the integration of rules and statistics presented, which takes advantage of a deep knowledge of the underlying rule-based system, especially in the way linguistic resources are used by the rule-based system when segmenting the source-language sentences; (*ii*) a complete analysis of the hybrid system, including automatic and manual evaluation, different corpora sizes, domains and language pairs, and a comparison to another popular hybrid MT system (Eisele, Federmann, Uszkoreit, Saint-Amand, Kay, Jellinghaus, Hunsicker, Herrmann, & Chen, 2008).

Cross-Lingual Bridges with Models of Lexical Borrowing by *Tsvetkov and Dyer* introduces a hybrid model of lexical borrowing, which is demonstrated in machine translation to alleviate the problem of lexical coverage in low-resourced languages. The authors start from the hypothesis that *all languages borrow terms from other languages at some point in their existence*. They propose a computational model of linguistic borrowing, intended to identify donor words in a resource-rich language given a loan word in a resource-poor language. The model develops a set of morpho-phonological transformations by combining linguistic constraints using optimality theory and machine learning to score loanword candidates. This model consists of three parts: (*i*) conversion of orthographic word forms to pronunciations,(*ii*) generation of loan word pronunciation candidates, and (*iii*) ranking of generated candidates using optimality-theoretic constraints. The first two steps are rule-based while the third is learned from data.

The lexical borrowing model is applied to Swahili-English, Maltese-English, and Romanian-English MT. The authors leverage the model in an indirect way. For instance, for improving the resource-poor Swahili-English MT system, they identify translation candidates for out-of-vocabulary (OOV) Swahili words borrowed from Arabic, using an Arabic-to-Swahili borrowing model and a resource-rich Arabic-English MT system. The experimental results show that this approach effectively reduces the impact of OOV source words and improves translation quality significantly.

If one considers lexical borrowing as a different task than transliteration and cognate identification then this could be the first computational model of lexical borrowing used in a downstream natural language processing application.

2.2 Domain and Language Adaptation for Sentiment

There is a rapidly growing repository of user-generated, subjective texts on the Internet in the form of blogs, social networks, information channels and consumer sites expressing opinions on various issues, sentiment towards products, and services and personal perspectives about events.

Sentiment analysis is the task of analyzing opinions, sentiments or emotions expressed towards entities such as products, services, organizations, issues, and the various attributes of these entities (Liu, 2012). The two main sentiment analysis approaches presented in the literature are a machine learning approach (mostly supervised learning) and an opinion-lexicon-based approach, based on rules.

When necessary resources—the training data, the subjective text to be analyzed and the analysis outcome—are not all available in the required language, cross-language sentiment analysis methods are needed to bootstrap a system. The main cross-language sentiment analysis approaches described in the literature are via lexicon transfer, via corpus transfer, via test translation and via joint classification. In the lexicon transfer approach, a source sentiment lexicon is transferred into the target language and a lexicon-based classifier is built in the target language (Mihalcea, Banea, & Wiebe, 2007). The corpus transfer approach consists of transferring a source training corpus into the target language and building a corpus-based classifier in the target language (Banea, Mihalcea, & Wiebe, 2008). In the test translation approach, test sentences from the target language are translated into the source language and they are classified using a source language classifier (Bautin, Vijayarenu, & Skiena, 2008). Work on joint classification includes co-training (Wan, 2009), joint learning (Lu, Tan, Cardie, & K. Tsou, 2011) or structural correspondence learning (Wei & Pal, 2010; Prettenhofer & Stein, 2010).

Some authors have already studied the impact on automatic sentiment analysis of transferring lexicons or corpora via machine translation (Mihalcea et al., 2007; Banea et al., 2008).

How Translation Alters Sentiment by *Mohammad, Salameh and Kiritchenko* goes a step further in the analysis in two respects. First, the authors conduct a systematic evaluation of the impact of automatic and manual translation on both automatic and manual sentiment analysis. Second, the authors perform a qualitative and quantitative analysis to understand the reasons for the obtained results. In summary, this paper provides a deeper understanding of how sentiments are altered in common cross-language settings.

The experiments are performed using an Arabic sentiment analysis system and an English–Arabic machine translation system, both showing state-of-the-art performance. The authors first show that automatic sentiment analysis of English translations (even coming from MT) can achieve competitive results. Interestingly, they also show that automatic sentiment analysis of automatic translations outperforms the manual sentiment annotation of the automatically translated text. The qualitative and quantitative analysis of the results also reveals interesting facts. For example, sentiment expressions are often mistranslated into neutral expressions. The automatic sentiment analysis system can recover from consistent translation errors by learning true sentiments from mistranslated words. Some common causes of translation failing to preserve sentiments are sarcasm, metaphoric expressions, and incorrect word-reordering.

Distributional Correspondence Indexing for Cross-Lingual and Cross-Domain Sentiment Classification by *Esuli, Moreo and Sebastiani* proposes a novel domain adaptation method, also evaluated on language adaptation. This paper explores a more general and complex formulation of the domain adaptation problem that combines the *cross-domain* and *cross-language* settings. The proposed adaptation method, called Distributional Correspondence Indexing, is inspired by Structural Correspondence Learning but follows a different, simpler approach, with a more direct application of the distributional hypothesis. This approach assumes that terms across domains and/or languages show similar distributional properties relative to a small set of "pivot" terms, which behave similarly across domains/languages. The authors show that this approach outperforms existing methods for cross-domain/language sentiment classification, at a lower computational cost.

Since digital documents in an increasing variety of topics and languages are produced, often with a need to be processed immediately, better solutions to tackle the bottleneck of scarcity of training data are of increasing importance. The idea of leveraging resources from a language to learn a classifier in another language is similar to transfer learning in domain adaptation. The contribution of this paper goes in the direction of unifying domain and language adaptation in the same framework.

2.3 Lexical Resources

More and more natural language applications such as question-answering, sentiment analysis, or document classification, to mention but a few, demand lexical knowledge in different natural languages, termed here cross-language lexical resources, and also well-known as multilingual lexical resources. These range from non-structured resources such as parallel corpora (EU JRC-Acquis Corpus, see Steinberger, Pouliquen, Widiger, Ignat, Erjavec, Tufis, & Varga, 2006), glossaries (e.g., IFLA Multilingual Glossary for Art Librarians, see Libraries, 1996) or machine-readable dictionaries (Oxford online dictionaries[2]), to more structured resources such as terminological databases (IATE[3]), thesauri (AGROVOC [4]), lexicons (EuroWordNet, see Vossen, 1998; MultiWordNet, see Pianta, Bentivogli, & Girardi, 2002), or ontologies (e.g., EUROVOC in SKOS, see Smedt & Vatanat, 2009; or FAO geopolitical ontology, see Kim, Iglesias-Sucasas, & Viollier, 2013).

The creation of such multilingual resources involves manual and costly processes, which is why many approaches pursue the automation of several steps in the development process. Without the aim of being exhaustive, we briefly describe typical approaches or methods that result in multilingual lexical resources. (*i*) The well-known family of multilingual wordnets developed around the Princeton English WordNet (Fellbaum, 1998). Basically, two approaches are followed: (*a*) a merging approach in which wordnets are created separately and mapped afterwards to the English WordNet, with the difficulties that the mapping task involves; and (*b*) an expansion approach in which the English WordNet is translated into the corresponding target languages, facilitating the subsequent mapping task. (*ii*) Online-collaborative resources, as for example, Wiktionary[5]. They take advantage of the wisdom of the crowd and also of Internet bots that automatically generate entries, as well as of algorithms that import lexical information from machine readable dictionaries. Similarly, Wikipedia and the whole family of Wikimedia projects have been created in a collaborative manner and constitute *de facto* multilingual resources thanks to the hyperlinks among entries in different languages. Other resources built in a similar way or that take advantage of the structured information contained in them include YAGO (Mahdisoltani, Biega, & Suchanek, 2015) and BabeLNet (Navigli & Ponzetto, 2012). (*iii*) Mono- and multilingual content and linguistic datasets exposed and linked according to the Linked Data paradigm[6], a set of best practices for publishing structured data and linking it to other datasets. Some

2. http://www.oxforddictionaries.com
3. http://iate.europa.eu
4. http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
5. https://www.wiktionary.org
6. http://linkeddata.org

of the resources published as linked data are multilingual and others are monolingual, but once linked to other datasets in the same or close domains, they become a multilingual graph of navigable data. The mapping or linking step is crucial in this approach. (*iv*) Translation resources can also be considered as a subtype of lexical resources and are also used for the purposes of obtaining multilingual data, be it as a source for translations, or as a means for mapping or linking monolingual resources (for more details on machine translation see section 2.1).

In this sense, it has been sufficiently demonstrated that quality and coverage of such lexical and translation resources are vital for applications built on top of them, and that the performance of those applications depends directly on them. The paper included in this special issue not only describes and analyzes some of these resources in detail, but also evaluates their coverage and correctness in the context of ontology mapping.

Effectiveness of Automatic Translations for Cross-lingual Ontology Mapping by *Abu Helou and Palmonari* presents a large-scale study on the effectiveness of several lexical and translation resources for the purpose of obtaining candidate matches between ontology concepts lexicalized in different languages. Many works on cross-language ontology mapping rely on multilingual or translation resources to obtain translation candidates for concept lexicalization in the source ontology, which subsequently support the selection of potential matches in the target ontology. This paper evaluates a machine translation service, GoogleTranslate[7], and a multilingual encyclopedic dictionary and semantic network, BabelNet[8], for correctness and coverage of the suggested translations and mapping selection capabilities (word-disambiguation). The evaluation is based on wordnets in different natural languages (Arabic, Italian, Slovene and Spanish) and manual alignments provided for each wordnet and the English WordNet. They perform three experiments taking into account various types of lexical units (monosemous words, polysemous words, one-word units or multiple-word units), and define specific measures (translation correctness, word sense coverage, synset coverage and synonym coverage) that serve to better evaluate the quality of translations. The results of the experiments provide insights into the coverage and correctness that these resources provide, the effect of combining the results from both resources, the impact of taking into account translation directionality, and the differences in word categories between mapped lexicalizations, among others.

The results of this paper can be directly applied to improve the cross-language ontology mapping task, but can also contribute to speed up the development of multilingual lexical resources, in general, or help in the building of a truly multilingual Linked Open Data Cloud, in particular.

2.4 Cross-Language Dependency Parsing

A natural language application that relies on language processing tools such as part-of-speech taggers and parsers is confronted with a significant challenge of scaling to new languages, when such languages may not have the same set of tools. Building such tools in a desired language requires manual annotation of sufficient amounts of texts that machine learning programs can be trained on. Annotation efforts have been undertaken and

7. https://translate.google.com
8. http://babelnet.org

invaluable resources of varying amounts of texts have been created during the past decades for certain languages, for example, Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993), French Treebank (Abeillé, 2003), NEGRA Treebank (Skut, Brants, & Uszkoreit, 1998), Prague Dependency Treebank (Hajič, Böhmová, Hajičová, & Vidová-Hladká, 2000). However, the task of such annotation effort is time-consuming, expensive and requires highly skilled personnel.

In the late nineties, with the availability of texts in a pair of languages as realized in parallel text corpora, researchers (Bangalore, 1998; Yarowsky, Ngai, & Wicentowski, 2001) explored the idea of transferring annotations from one language in the pair to the second language in the pair in order to rapidly create an annotated resource to train language processing tools for the second language. This line of research has been followed by a number of researchers for projecting a variety of annotations between language corpora and using the annotated corpus to bootstrap language processing tools for the target language. Of particular interest is the projection of annotations for part-of-speech tags, phrase structure annotations, and dependency structure annotations.

Synthetic Treebanking for Cross-lingual Dependency Parsing by *Tiedemann and Agić* presents a detailed discussion of the options to transplant dependency trees from one language to another in order to train dependency parsers in the target language. The paper introduces two options of bootstrapping a dependency parser for a language: a model transfer approach where a dependency parsing model is transferred to the new language, and an annotation transfer approach; the paper then advocates the annotation transfer approach through the creation of synthetic treebanks. The paper provides a comprehensive analysis of the various options for creating such a synthetic treebank using a parallel corpus including: (*i*) projecting a parser's output of the source text onto the target text and (*ii*) translating an existing high quality treebank in a source language to a target language. The central idea in creating synthetic treebanks involves the use of statistical machine translation models, both phrase-based and syntax-based, in order to translate the texts of the source language treebanks into the target language, and then project the source language structures to the translated target language sentence mediated by the word alignment information produced from the translation process. The challenges of reconciling the dependency structures when the lexical translations are richer than one-to-one are discussed at length. The paper reports extensive parsing accuracy results from parsers trained on projected treebanks for a large set of language pairs, and studies the correlation between the quality of translation model and the quality of the target language parser.

2.5 Cross-Language Information Retrieval

With the increase in number of webpages in multiple languages, a search query on the Web needs to retrieve webpages authored in languages other than the language of the query. Cross-language information retrieval is a technology at the intersection of machine translation and information retrieval that addresses this challenge.

Cross-language information retrieval has employed different strategies for matching a query with a set of multilingual documents: cognate-based matching (Montalvo, Martínez, Casillas, & Fresno, 2007), matching by query and document translation, and matching by mapping to an interlingua (Banchs & Costa-jussà, 2013). The most popular of these approaches is query translation and it can be addressed by either employing bilingual dictionaries (Hedlund, Airio, Keskustalo, Lehtokangas, Pirkola, & Jarvelin, 2004) or using machine translation (Kishida, 2008). Approaches that combine both techniques can be found in the work of in the work of Zhang, Jones, and Zhang (2008), for example. The quality of query translation can be indirectly observed in the final retrieval results. Kishida (2008) shows with a regressive model that both ease of search of a given query and translation quality can explain about 60% of the variation in the performance. Kettunen (2009) shows that for long topics, the correlations between achieved retrieval results and machine translation metrics are high (almost 90%) and for short topics the correlation is lower but still clear (almost 60%). Cross-language video retrieval, another related cross-language information retrieval task, involves automatic speech recognition.

Utilisation of Metadata Fields and Query Expansion in Cross-Lingual Search of User-Generated Internet Video by *Khwileh, Jones and Ganguly* presents one of the first use-cases for cross-language video retrieval for social media content. The paper focuses on all the challenges associated with user generated informal content (i.e., noise, sparseness of metadata, content with very different lengths, informal language register, etc.) rather than on professionally produced content. Noise and errors propagate through each step of the processing: from speech recognition to automatic translation and query expansion. The authors use the query translation approach to bridge the vocabulary gap between the user's query and the relevant content in a video application. Automatic translation is done using Google Translate and retrieval and expansion is done with the Divergence From Randomness IR model. They explore the effectiveness of three different sources of information: transcripts from automatic speech recognition, video titles, and descriptions. Among the three sources, leveraging video titles improves retrieval performance in their experiments. In addition, the authors propose an adaptive query expansion technique that automatically selects the most reliable source for expansion based on a well established query performance prediction technique. Results show that this approach is more robust for this particular setting.

2.6 Cross-Language Knowledge Representation

Knowledge representation systems aim to formalize representations of the world, or of a certain domain of knowledge, in such a way that they are interpretable by computers. This is achieved by identifying the domain concepts and the relations that exist among them, and by representing all this information in a formal framework, e.g., using the Description Logic formalism. Knowledge representation systems are intended to be language-independent meaning representations. However, different representations of the same domain of knowledge can co-exist, since those who design a certain knowledge representation may have a particular vision of the world, specific interests, or do it with a certain application in mind, among other factors.

One of the difficulties of cross-language knowledge representation regards the conceptual differences that can be observed across languages and cultures. Indeed, certain representations are prone to reflect cultural particularities that are not shared or understood in the same way by other cultural systems. This involves the existence of certain concepts that do not exist in other knowledge systems, or are not relevant to them, or are at different granularity levels in the representation of concepts (Espinoza, Montiel-Ponsoda, & Gómez-Pérez, 2009). For all these reasons, cross-language knowledge representation is a challenge in the current multilingual Web. Two main approaches have been followed to obtain knowledge representation systems that support several languages: (*i*) inclusion of lexicalizations in several natural languages to describe the concepts and relations formalized in a certain knowledge representation system; (*ii*) existence of several knowledge representation systems whose concepts and relations are expressed for a different natural language, which are then linked or mapped to establish correspondences or equivalences between them. The former approach is commonly applied in internationalized or standardized domains, whereas the latter is typical in culturally-influenced domains, as termed by Cimiano, Montiel-Ponsoda, Buitelaar, Espinoza, and Gmez-Prez (2010).

In our globalized and interconnected world, cross-language information access is of increasing importance. While several approaches to cross-language document similarity have been reported, including machine translation, probabilistic topic models, classification or matrix factorization, there is little previous work on the task of linking documents across languages that refer to the same events (Pouliquen, Steinberger, Ignat, Ksper, & Temnikova, 2004; Pouliquen, Steinberger, & Deguernel, 2008; Leban, Fortuna, Brank, & Grobelnik, 2014). This is actually a difficult and computationally expensive task, especially when many language pairs are involved, and only a very small number of real-life working systems performing this task exist. One example of an existing service providing cross-language cluster linking is the European Media Monitor (EMM) (Pouliquen et al., 2008). This special issue includes a contribution tackling this task, in which topic modeling is used to represent the knowledge expressed in documents, and the linking task is based on similarity-based and entity-related features.

News Across Languages - Cross-Lingual Document Similarity and Event Tracking by *Rupnik, Muhic, Leban, Skraba, Fortuna and Grobelnik* addresses the problem of event tracking in a large multilingual stream, and, more specifically, how to link collections of articles in different languages which refer to the same event. The authors consider major languages and also less-resourced languages. The approach is based on representations of documents analogous to multilingual topics, which are valid over multiple languages. These representations are learned using Wikipedia as a training corpus. They are then used to compute cross-language similarities between documents regardless of language. The posterior cross-language cluster linking is performed in two steps. First, to speed-up the process, the similarity function is used to identify a small set of potential linking candidates for each cluster. Then, the final decision is taken based on a supervised classification model whose features include similarity-based and entity-related features. In a comprehensive experimental study, the authors show canonical correlation analysis to be the best-performing method to compute multilingual similarities. Moreover, they show that similarity-based features can greatly benefit from additional semantic extraction-based features.

Acknowledgements

The authors want to thank Dan Roth, Mark Sammons and an anonymous reviewer for their useful comments and suggestions on previous versions of this document. This work has been supported by the 7th Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951), the Intra-European Fellowship CrossLingMind-2011-300828 and the project LIDER (610782); the European Regional Development Fund (ERDF/FEDER); and the Spanish Ministerio de Economía y Competitividad through the SpeechTech4All project (TEC2012-38939-C03-02) and the project "4V: volumen, velocidad, variedad y validez en la gestión innovadora de datos" (TIN2013-46238-C4-2-R).

References

Abeillé, A. (2003). *Treebanks: Building and Using Parsed Corpora*. Springer.

Armentano-Oller, C., & Forcada, M. L. (2006). Open-source machine translation between small languages: Catalan and aranese occitan. In *Workshop on Strategies for developing machine translation for minority languages*, pp. 51–54.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, *abs/1409.0473*.

Banchs, R., & Costa-jussà, M. R. (2013). Cross-Language Document Retrieval by using Non-linear Semantic Mapping. *Applied Artificial Intelligence Journal*, *27*(9), 781–802.

Banea, C., Mihalcea, R., & Wiebe, J. (2008). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In *Proceedings of the International Conference on Linguistic Resources and Evaluation (LREC)*, pp. 2764–2767, Marrakech, Morocco.

Bangalore, S. (1998). Transplanting Supertags from English to Spanish. In *Proceedings of the TAG+4 Workshop*.

Bautin, M., Vijayarenu, L., & Skiena, S. (2008). International Sentiment Analysis for News and Blogs. In *Proc. of the International Conference on Weblogs and Social Media*, pp. 19–26, Seattle, U.S.A.

Bener, A., Minku, L., & Turhan, B. (Eds.). (2014). *PROMISE '14: Proceedings of the 10th International Conference on Predictive Models in Software Engineering*, New York, NY, USA. ACM.

Bikel, D., & Zitouni, I. (2012). *Multilingual Natural Language Processing Applications: From Theory to Practice*. IBM Press.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., & Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 1–46, Lisbon, Portugal.

Chiang, D. (2007). Hierarchical Phrase-Based Translation. *Computational Linguistics*, *33*(2), 201–228.

Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., & Gómez-Pérez, A. (2010). A Note on Ontology Localization. *Journal of Applied Ontology*, *5(2)*, 127–137.

Costa-jussà, M. R. (2015). How Much Hybridization Does Machine Translation Need?. *Journal of the American Society for Information Technology (JASIST)*, *6*(10), 2160–2165.

Costa-jussà, M. R., Banchs, R., Rapp, R., Lambert, P., Eberle, K., & Babych, B. (2013). Workshop on hybrid approaches to translation: Overview and developments. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pp. 1–6, Sofia, Bulgaria.

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., & Chen, Y. (2008). Hybrid Architectures for Multi-Engine Machine Translation. In *Proceedings of Translating and the Computer 30*. ASLIB/IMI, ASLIB.

Espinoza, M., Montiel-Ponsoda, E., & Gómez-Pérez, A. (2009). Ontology Localization. In *Proceedings of the 5th International Conference on Knowledge Capture (KCAP09)*, pp. 33–40.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Forner, P., Moller, H., Paredes, R., Rosso, P., & Stein, B. (2013). *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. Springer.

Gracia, J., MacCrae, J., & Vulcu, G. (Eds.). (2015). *Proceedings of the Fourth Workshop on the Multilingual Semantic Web*. CEUR.

Hajič, J., Böhmová, A., Hajičová, E., & Vidová-Hladká, B. (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Abeillé, A. (Ed.), *Treebanks: Building and Using Parsed Corpora*, pp. 103–127. Amsterdam:Kluwer.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., & Jarvelin, K. (2004). Dictionary-based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000-2002. *Information Retrieval*, *7*(1), 99–119.

Hutchins, W. J., & Sommers, H. L. (1992). *An Introduction to Machine Translation*, Vol. 362. Academic Press, New York.

Kettunen, K. (2009). Choosing the Best MT Programs for CLIR Purposes—Can MT Metrics Be Helpful?. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pp. 706–712.

Kim, S., Iglesias-Sucasas, M., & Viollier, V. (2013). The FAO Geopolitical Ontology: A Reference for country-Based Information. *Journal of Agricultural and Food Information*, *14*(1).

Kishida, K. (2008). Prediction of Performance of Cross-language Information Retrieval Using Automatic Evaluation of Translation. *Library and Information Science Research*, *30*(2), 138–144.

Koehn, P., Och, F., & Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*, pp. 48–54.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pp. 177–180.

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014). Event registry: Learning about world events from news. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion (WWW Companion'14)*, pp. 107–110, Seoul, Korea.

Libraries, I. S. o. A. (Ed.). (1996). *Multilingual Glossary for Art Librarians: English with Indexes in Dutch, French, German, Italian, Spanish and Swedish*. De Gruyter.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Lu, B., Tan, C., Cardie, C., & K. Tsou, B. (2011). Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 320–330, Portland, Oregon, USA.

Mahdisoltani, F., Biega, J., & Suchanek, F. M. (2015). YAGO3: A Knowledge Base from Multilingual Wikipedias. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR 2015)*.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning Multilingual Subjective Language via Cross-Lingual Projections. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 976–983, Prague, Czech Republic.

Montalvo, S., Martínez, R., Casillas, A., & Fresno, V. (2007). Multilingual News Clustering: Feature Translation vs. Identification of Cognate Named Entities. *Pattern Recognition Letters*, *28*(16), 2305–2311.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, *193*, 217–250.

Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the Frist International Conference on Global WordNet*.

Pouliquen, B., Steinberger, R., & Deguernel, O. (2008). Story Tracking: Linking Similar News Over Time and Across Languages. In *Proceedings of the COLING 2008 Workshop on Multi-source Multilingual Information Extraction and Summarization*, pp. 49–56, Manchester, UK.

Pouliquen, B., Steinberger, R., Ignat, C., Ksper, E., & Temnikova, I. (2004). Multilingual and cross-lingual news topic tracking. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 959–965, Geneva, Switzerland. COLING.

Prettenhofer, P., & Stein, B. (2010). Cross-Language Text Classification Using Structural Correspondence Learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1118–1127, Uppsala, Sweden.

Quirk, C., Menezes, A., & Cherry, C. (2005). Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 271–279.

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., & Pérez-Ortiz, J. A. (2011). The Universitat d'Alacant hybrid machine translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 457–463, Edinburgh, Scotland.

Sánchez-Martínez, F., & Forcada, M. L. (2009). Inferring Shallow-Transfer Machine Translation Rules from Small Parallel Corpora. *Journal of Artificial Intelligence Research*, *34*, 605–635.

Skut, W., Brants, T., & Uszkoreit, H. (1998). A Linguistically Interpreted Corpus of German Newspaper Text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrucken, Germany.

Smedt, J. D., & Vatanat, B. (2009). https://lists.w3.org/archives/public/public-esw-thes/2010feb/att-0023/ontology.html..

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy.

Suchanek, F., Riedel, S., Singh, S., & Pratim Talukdar, P. (Eds.)., (2012). *AKBC-WEKEX '12: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vossen, P. (Ed.). (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Wan, X. (2009). Co-Training for Cross-Lingual Sentiment Classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 235–243, Singapore.

Wei, B., & Pal, C. (2010). Cross Lingual Adaptation: An Experiment on Sentiment Classifications. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 258–262, Uppsala, Sweden.

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pp. 1–8, San Diego, CA, USA.

Zhang, Y., Jones, G. J., & Zhang, K. (2008). Dublin City University at CLEF 2007: Cross-Language Speech Retrieval Experiments. *Lecture Notes In Computer Science*, 703–711.