Semantic Indexing of Multilingual Corpora and its Application on the History Domain

Alessandro Raganato[1], **Jose Camacho-Collados**[1], **Antonio Raganato**[2] and **Yunseo Joung**
Department of Computer Science[1], Department of Political Sciences[2]
Sapienza University of Rome
{raganato,collados}@di.uniroma1.it
{raganatoantonio,louisejoung}@gmail.com

Abstract

The increasing amount of multilingual text collections available in different domains makes its automatic processing essential for the development of a given field. However, standard processing techniques based on statistical clues and keyword searches have clear limitations. Instead, we propose a knowledge-based processing pipeline which overcomes most of the limitations of these techniques. This, in turn, enables direct comparison across texts in different languages without the need of translation. In this paper we show the potential of this approach for semantically indexing multilingual text collections in the history domain. In our experiments we used a version of the Bible translated in four different languages, evaluating the precision of our semantic indexing pipeline and showing its reliability on the cross-lingual text retrieval task.

1 Introduction

In recent years there has been a growing interest in automatically processing historical corpora due to the increasing number of available text collections in the field (Dekkers et al., 2009). However, few software applications for non-expert users have been developed for processing and indexing historical texts, and these applications are in the main based on statistical processing techniques only (Piotrowski, 2012). Even though these techniques have been and are currently widely used, they have clear limitations. First, standard statistical processing techniques based on keywords do not handle the inherent ambiguity within language. Second, occurrences of the same concept/event/entity are often referred to via different lexicalizations (e.g. *Louis XIV*, *Louis the Great* and *Sun King*), which are not captured by keyword-based text retrieval techniques. Finally, these approaches are bound to remain monolingual by nature, limiting their applicability to multilingual corpora, which is growing in interest over the years (Johansson, 2007). There have been recent approaches to automatically link cultural heritage items from text corpora to knowledge bases (Brugman et al., 2008; Fernando and Stevenson, 2012; Hall et al., 2012; Efremova et al., 2014; Poelitz and Bartz, 2014) but without going beyond the monolingual level. In fact, to date most approaches towards the accessibility of cultural heritage content in multiple languages have focused on the generation of natural language content through knowledge bases or via the Semantic Web (Davies, 2009; Dannélls et al., 2013).

Instead, we propose a knowledge-based pipeline for automatically processing multilingual corpora which overcomes all previously mentioned limitations by going beyond standard statistical techniques and keyword-based queries. Our approach is based on the disambiguation of text corpora through a knowledge base. Disambiguation is then exploited for semantically indexing multilingual text collections by associating each concept/entity with a unique identifier independent on the language and the surface form. This in turn enables direct applications across languages such as cross-lingual text retrieval, and opens up new lines of research to study cross-cultural differences from multilingual text corpora (Gutiérrez et al., 2016).

Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH),
pages 140–147, Osaka, Japan, December 11-17 2016.

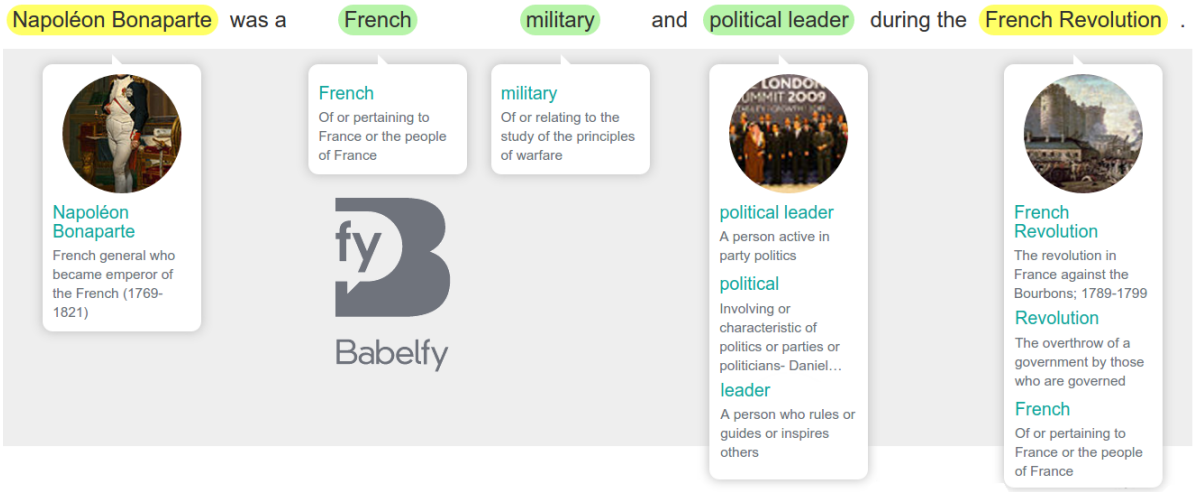

Figure 1: Sample output disambiguation.

2 Methodology

In this section we explain our pipeline for the semantic processing of multilingual corpora. For the semantic processing we rely on BabelNet (Navigli and Ponzetto, 2012), a large multilingual encyclopedic dictionary and semantic network. BabelNet[1] integrates various resources such as WordNet (Miller, 1995), Open Multilingual WordNet (Bond and Foster, 2013), Wikipedia, OmegaWiki, Wiktionary and Wikidata, among others. All the aforementioned resources are merged into a very large lexical resource in which equivalent concepts and entities are aggregated from the different resources in a unique instance, called BabelNet synset. Each synset contains all the synonyms and definitions harvested from the respective resources in a range of different languages. In fact, BabelNet includes 271 languages and has already shown its potential in various multilingual and cross-lingual Natural Language Processing applications (Moro et al., 2014; Camacho-Collados et al., 2015; Camacho-Collados et al., 2016b). We propose to use this knowledge base to semantically index large collections of multilingual texts. Our methodology is divided in two main steps: (1) corpus preprocessing including disambiguation and entity linking (Section 2.1) and (2) semantic indexing (Section 2.2).

2.1 Disambiguation and Entity Linking

The goal of this step is to associate each content word[2] with a unique unambiguous identifier (i.e., a BabelNet synset). First, texts are preprocessed (tokenized, Part-Of-Speech tagged and lemmatized) using Stanford CoreNLP (Manning et al., 2014) and TreeTagger (Schmid, 1994) on the languages for which these tools are available. For the remaining languages we rely on the multilingual preprocessing tools integrated in Babelfy. Since the disambiguation is targeted to historical texts, we include a list of stopwords belonging to the archaic form of a given language for the languages for which this list is available. For example, for English we used a list[3] including archaic expressions such as *thou* or *ye*. These stopwords are therefore not taken into account in the disambiguation process.

Then, preprocessed texts[4] are disambiguated using Babelfy (Moro et al., 2014), a state-of-the-art knowledge-based Word Sense Disambiguation and Entity Linking system based on BabelNet. Babelfy[5] exploits a densest subgraph heuristic for selecting high-coherence semantic interpretations of the input text and has been shown to perform on par with supervised systems on both Word Sense Disambiguation

[1] http://babelnet.org
[2] Multiwords are also considered on the disambiguation.
[3] http://bryanbumgardner.com/elizabethan-stop-words-for-nlp/
[4] As mentioned earlier, for the languages not covered by Stanford CoreNLP and TreeTagger we directly rely on the Babelfy pipeline.
[5] http://babelfy.org

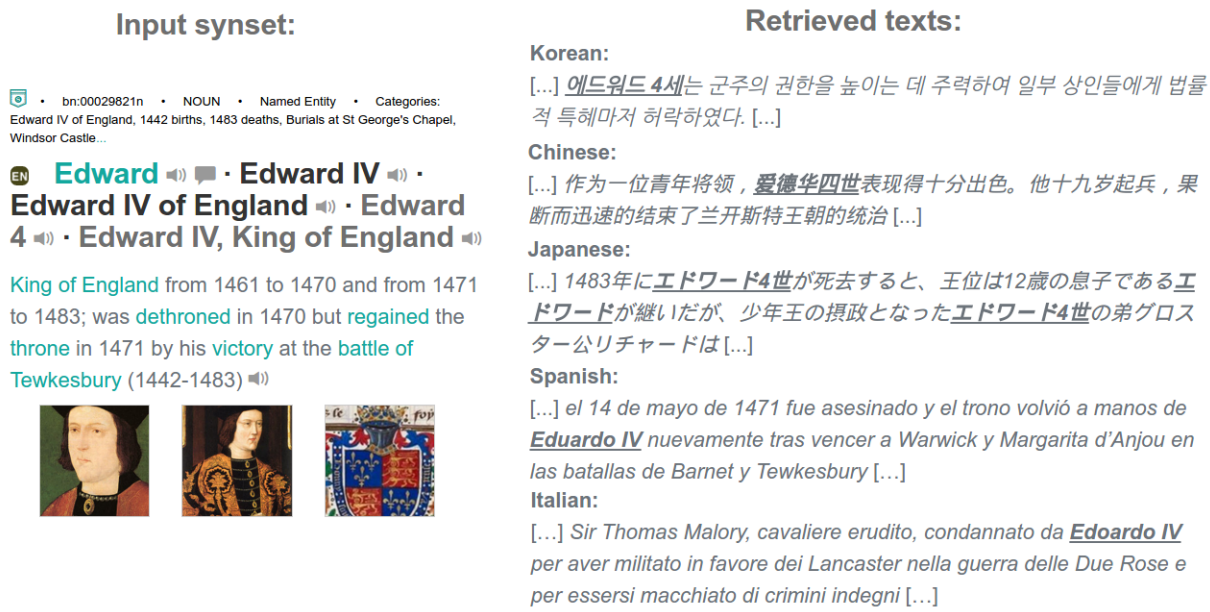

Figure 2: Semantic indexing of multilingual corpora: sample retrieved texts given *Edward IV* as input.

and Entity Linking tasks. Figure 1 shows a sample output disambiguation of a sentence as given by Babelfy.

2.2 Semantic indexing

Finally, for indexing a given text collection we directly use the output provided on the disambiguation step. Given a certain BabelNet synset or an instance from Wikipedia (recall from Section 2 that BabelNet is a multilingual resource containing Wikipedia among other resources), our system provides all the texts in a given corpus containing that instance. This is particularly interesting when the corpus is composed by texts in different languages as it directly benefits from the multilingual disambiguation performed in the previous section. Instead of translating a given concept or entity in different languages, our model is able to retrieve all the texts in which the given concept or entity occurs, irrespective of the text language. For instance, given a corpus of texts in different languages about the Late Middle Ages, our system would automatically retrieve all the texts in which the king of England *Edward IV* occurs (see Figure 2). This may be especially useful for carrying out a research on a specific person/event, as it differs from usual searches based on keywords which are focused on a single language and do not deal with ambiguity and synonymy.

3 Cross-lingual text retrieval

One of the most straightforward applications from the semantic indexing of corpora is cross-lingual text retrieval. The task of cross-lingual text retrieval consists of, given an input query, retrieving the texts which are more relevant to the input query. In this case, the user introduces a text as input (included in the corpus or not) and as an output our system retrieves the n most similar texts within the corpus, which may be written in a different language from the language of the input text. This application may be particularly useful to retrieve texts referring to the same period of history in large collections of corpora in different languages. Unlike usual monolingual retrieval systems based on word overlapping, our semantic pipeline can seamlessly retrieve texts in any given language thank to the disambiguation step (see Section 2.1).

Our approach to the cross-lingual retrieval of texts fully relies on the disambiguation performed for semantically indexing text collections. Each text is associated with the set of its disambiguated instances. Then, we simply use the Jaccard similarity coefficient for sets (Jaccard, 1901) to measure the

similarity between different texts. Since disambiguated instances (i.e., multilingual BabelNet synsets) are comparable across languages, no translation is needed to measure the similarity. In Section 4.2 we show the effectiveness of this approach for the task, performing on par or better than models requiring a pre-translation step.

4 Evaluation

We perform an evaluation to test the disambiguation quality of our multilingual semantic processing pipeline (Section 4.1) and the cross-lingual text retrieval application (Section 4.2). For both evaluations we use the same reference corpus, which is the Bible[6] translated into four different languages: English, Spanish, French, and Russian. Each language version consists of 1189 chapters of different sizes, ranging from 21 to 2423 words (588 words on average).

4.1 Disambiguation

In order to measure the disambiguation quality of Babelfy in the history domain, we manually annotated two chapters of the Bible for English and Spanish[7]. Table 1 shows the precision of our system and the Most Common Sense[8] (MCS) baseline in the evaluation set. Babelfy outperforms the MCS baseline in both languages, obtaining an overall precision of 68.8% for English and 58.8% for Spanish. Not surprisingly, the history domain is hard for a standard disambiguation system, which has been shown to perform above 70% in news corpora (Navigli et al., 2013). However, for nouns, which are the items our pipeline is especially targeted for, our system achieves considerably better results (63.4% for Spanish and 74.2% for English). The results of an open-domain disambiguation system clearly improving over the MCS baseline are indeed encouraging towards the development of a domain-specific disambiguation system. As future work we plan to adapt the disambiguation pipeline to the history domain by both refining the sense inventory and training on domain-specific corpora.

		English	**Spanish**
All	Our	68.8	58.8
	MCS	51.1	44.0
Nouns	Our	74.2	63.4
	MCS	58.7	47.8

Table 1: Precision (%) of Babelfy after preprocessing and the MCS baseline in the Bible.

4.2 Cross-lingual text retrieval

In this section we evaluate the effectiveness of our cross-lingual text retrieval pipeline (Section 3). The experimental setup is described in Section 4.2.1 and the results are presented in Section 4.2.2.

4.2.1 Experimental setup

Task description. Given a chapter of the Bible in one language (i.e., input language), the task consists of retrieving the same chapter in another language (i.e., output language) among the 1189 possible chapters[9]. Formally, given a chapter of the Bible in the input language, the system calculates the similarity between the given chapter and all the chapters in the output language. The chapter of the output language obtaining the highest similarity score is selected as retrieved chapter for the system. This task is intended to test the cross-lingual text retrieval application proposed in Section 3, which is based in the semantic indexing presented in Section 2.2.

[6] homepages.inf.ed.ac.uk/s0787820/bible/

[7] We release this sense-annotated evaluation corpus of 594 annotations for the research community at our website.

[8] MCS has traditionally been a hard baseline to beat for automatic disambiguation systems (Navigli, 2009).

[9] Although the chapters in the Bible have not been translated literally from sentence to sentence (some chapters were rewritten differently for certain languages), the Bible may be viewed as a reliable chapter-aligned comparable corpus for the evaluation.

Input Language	System	English	Spanish	French	Russian
English	Our	-	99.4	98.7	96.9
	MT+Jacc.	-	99.8	99.8	99.7
	MT+W2V	-	88.4	81.4	82.3
Spanish	Our	99.2	-	99.8	96.6
	MT+Jacc.	99.8	-	99.8	99.8
	MT+W2V	88.8	-	99.0	97.5
French	Our	98.6	99.7	-	95.2
	MT+Jacc.	99.7	99.9	-	99.9
	MT+W2V	83.0	99.2	-	96.0
Russian	Our	97.6	98.1	96.7	-
	MT+Jacc.	99.9	99.7	99.7	-
	MT+W2V	91.1	98.2	97.0	-

Table 2: Accuracy (%) of all comparison systems for the cross-lingual text retrieval task in the Bible.

Comparison systems. We include two baselines relying on monolingual text similarity measures after translation, using English as pivot language. This monolingual similarity measurement after translation is the most common approach in cross-lingual text similarity tasks (Agirre et al., 2016). For these baselines all the Bible chapters in languages other than English were automatically translated to English using the *Bing Translator* Machine Translation system[10], which covers the four languages considered in the evaluation. The first baseline system (**MT+Jacc.**) calculates the similarity between the content words of the output texts after translation by using the Jaccard index. The second baseline (**MT+W2V**) leverages word embeddings to calculate the similarity between the translated texts. The similarity measure consists of the cosine similarity between the average vector of the content word embeddings of both respective translated texts. This approach based on the centroid vector is often used in the literature to obtain representations of sentences and documents (Chen et al., 2014; Yu et al., 2014). As word embeddings we use the pre-trained Word2Vec (Mikolov et al., 2013) vectors trained on the Google News corpus[11].

4.2.2 Results and Discussion

Table 2 shows the accuracy[12] results of all comparison systems in the cross-lingual text retrieval task using the Bible as gold standard comparable corpus for four different languages: English, Spanish, French, and Russian. Given the current state of MT systems, the high results obtained by the translation-based system are not surprising. However, our simple system based on inherently imperfect disambiguation achieves comparable results to the baseline based on the lexical similarity measure after translation (MT+Jacc.) and improves considerably the results of the system based on word embeddings after translation (MT+W2V). This improvement over the system based on word embeddings may be due to two main factors. First, since the translation is carried out automatically, it may be prompt to errors. Second, even though word embeddings have already shown its potential in obtaining accurate semantic representations of lexical items, they may not be so accurate to model larger semantic units such as documents. In fact, word embeddings are in the main used in tasks which make use of the local context of words, e.g., dependency syntactic parsing (Weiss et al., 2015; Bansal et al., 2014), rather than in tasks requiring the global semantic representations of documents or paragraphs.

The results are especially meaningful considering that our system does not require a prior translation step between languages. In fact, obtaining and integrating reliable translation models for all pairs of languages is generally a heavily impractical task (Jones and Irvine, 2013). This is definitely an encouraging

[10] `https://www.bing.com/translator`

[11] `https://code.google.com/archive/p/word2vec/`

[12] Accuracy is computed as the number of times a system retrieves the same chapter in the output language divided by the total number of chapters (i.e., 1189).

```
<dataset language="EN" title="GEN">
<paragraph id="p.1">
    <text>
        In the beginning God created the heaven and the earth.
        And the earth was without form, and void; and darkness was upon the face of the deep.
        And the Spirit of God moved upon the face of the waters.
        ...
    </text>
    <annotations>
        <annotation source="MCS" anchor="beginning" bfScore="--" coherenceScore="--">bn:00009632n</annotation>
        <annotation source="BABELFY" anchor="God" bfScore="0.7620" coherenceScore="0.7913">bn:00040878n</annotation>
        <annotation source="MCS" anchor="created" bfScore="--" coherenceScore="--">bn:00086008v</annotation>
        <annotation source="BABELFY" anchor="earth" bfScore="0.8485" coherenceScore="0.8079">bn:00029424n</annotation>
        ...
    </annotations>
</paragraph>
...
```

Figure 3: XML snippet from the Book of Genesis.

result towards the use of multilingual lexical resources as a bridge to connect corpora from different languages.

5 Release

As a result of this work, we provide a tool for semantically indexing any given corpus[13] and release it at `http://wwwusers.di.uniroma1.it/~raganato/semantic-indexing`. The tool is intended for non-expert users, i.e., users that do not require any prior programming knowledge.

First, the input corpus is disambiguated (see Section 2.1) and is automatically stored in standard XML-formatted files, following the annotation format used in Camacho-Collados et al. (2016a). In our case an XML file is produced for each document and documents are disambiguated paragraph by paragraph by default. Figure 3 shows a sample XML output file for a portion of the Bible. Each file contains a list of `paragraph` tags. Each paragraph tag is composed by the original plain text and its sense annotations. The `annotation` tag refers to the sense annotations provided as a result of the disambiguation process. Each annotation includes its disambiguated BabelNet id, containing four attributes:

- `source`: this attribute indicates whether the disambiguation has been performed by Babelfy or if the system has back-off to the Most Common Sense (MCS) heuristic.
- `anchor`: this attribute corresponds to the surface form of the concept or entity as found within the paragraph.
- `bfScore`: this attribute corresponds to the Babelfy confidence score.
- `coherenceScore`: this attribute corresponds to the coherence score[14].

Finally, we provide a simple interface where users may introduce unambiguous BabelNet ids or Wikipedia pages to retrieve their occurrences in the whole corpus. A user may also introduce a word (or a multiword expression) as input. In this case the interface would ask the user to provide the desired sense among all the options. For instance, if the user introduces *Alexander* as input, the user will be required to select between *Alexander the Great* or *Czar Alexander III* among others.

The cross-lingual text retrieval application is additionally included in the provided interface. For this application the user gives a document/paragraph as input and the system will retrieve the closest documents/paragraphs as given by our pipeline (see Section 3).

[13]The input corpus may be given as a collection of simple raw text files. More details on the required input format are provided in the website.

[14]See Camacho-Collados et al. (2016a) or Babelfy API guide (`http://babelfy.org/guide`) for more information about these two scores.

6 Conclusion and Future Work

In this paper we presented a pipeline for processing historical corpora and showed its potential for semantically indexing multilingual corpora and for the cross-lingual text retrieval task. We provide an interface for non-expert users for semantically indexing any given multilingual corpus, including a demo based on the Bible already processed for the four languages included in the evaluation: English, Spanish, French and Russian. Note that even though in this paper we have only discussed the potential of using our pipeline for historical texts, our pipeline may be used for multilingual corpora coming from different domains as well.

As future work we aim at improving the disambiguation pipeline on historical corpora by refining the semantic network of BabelNet to the history domain. Additionally, we plan to apply our pipeline to study the role of various historical characters according to texts from different cultures written in different languages.

Acknowledgments

The authors gratefully acknowledge the support of the Sapienza Research Grant Avvio alla Ricerca 2015 No. 56.

We would also like to thank the anonymous reviewers for their helpful comments and Claudio Delli Bovi for helping us with the annotations and for his comments on the manuscript.

References

Eneko Agirre, Carmen Baneab, Daniel Cerd, Mona Diabe, Aitor Gonzalez-Agirrea, Rada Mihalceab, German Rigaua, Janyce Wiebef, and Basque Country Donostia. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. *Proceedings of SemEval*, pages 497–511.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *ACL (2)*, pages 809–815.

Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *ACL (1)*, pages 1352–1362.

Hennie Brugman, Véronique Malaisé, and Laura Hollink. 2008. A common multimedia annotation framework for cross linking cultural heritage digital collections. In *Proceedings of LREC*.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A Unified Multilingual Semantic Representation of Concepts. In *Proceedings of ACL*, pages 741–751, Beijing, China.

José Camacho-Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli. 2016a. A Large-Scale Multilingual Disambiguation of Glosses. In *Proceedings of LREC*, pages 1701–1708, Portoroz, Slovenia.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016b. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP*, pages 1025–1035, Doha, Qatar.

Dana Dannélls, Aarne Ranta, Ramona Enache, Mariana Damova, and Maria Mateva. 2013. Multilingual access to cultural heritage content on the semantic web. *LaTeCH 2013*, page 107.

Rob Davies. 2009. EuropeanaLocal–its role in improving access to Europes cultural heritage through the European digital library. In *Proceedings of IACH workshop at ECDL2009 (European Conference on Digital Libraries)*, Aarhus, Denmark.

Makx Dekkers, Stefan Gradmann, and Carlo Meghini. 2009. Europeana outline functional specification for development of an operational european digital library. *Europeana Thematic Network Deliverable*, 2.

Julia Efremova, Bijan Ranjbar-Sahraei, and Toon Calders. 2014. A hybrid disambiguation measure for inaccurate cultural heritage data. In *The 8th workshop on LaTeCH*, pages 47–55. Citeseer.

Samuel Fernando and Mark Stevenson. 2012. Adapting wikification to cultural heritage. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 101–106. Association for Computational Linguistics.

E. D. Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

Mark M Hall, Paul D Clough, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. 2012. Enabling the discovery of digital cultural heritage objects through wikipedia. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 94–100. Association for Computational Linguistics.

Paul Jaccard. 1901. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge.

Stig Johansson. 2007. Seeing through multilingual corpora. *Language and Computers*, 62(1):51–71.

Ruth Jones and Ann Irvine. 2013. The (un) faithful machine translator. *LaTeCH 2013*, page 96.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proceedings of SemEval 2013*, pages 222–231.

Roberto Navigli. 2009. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.

Christian Poelitz and Thomas Bartz. 2014. Enhancing the possibilities of corpus-based investigations: Word sense disambiguation on query results of large text corpora. *EACL 2014*, page 42.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49.

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. In *Proceedings of ACL*, page 323333.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. In *NIPS Deep Learning Workshop*.