Multilingual Word Sense Induction to Improve Web Search Result Clustering

Lorenzo Albano
University of Modena and
Reggio Emilia
41125 via P. Vivarelli 10
Modena, Italy

Domenico Beneventano
University of Modena and
Reggio Emilia
41125 via P. Vivarelli 10
Modena, Italy
name.surname@unimore.it

Sonia Bergamaschi
University of Modena and
Reggio Emilia
41125 via P. Vivarelli 10
Modena, Italy

ABSTRACT

In [12] a novel approach to Web search result clustering based on Word Sense Induction, i.e. the automatic discovery of word senses from raw text was presented; key to the proposed approach is the idea of, first, automatically inducing senses for the target query and, second, clustering the search results based on their semantic similarity to the word senses induced. In [1] we proposed an innovative Word Sense Induction method based on multilingual data; key to our approach was the idea that a multilingual context representation, where the context of the words is expanded by considering its translations in different languages, may improve the WSI results; the experiments showed a clear performance gain. In this paper we give some preliminary ideas to exploit our multilingual Word Sense Induction method to Web search result clustering.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing-Text analysis; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

General Terms

Algorithms, Experimentation

Keywords

Multilingual, Word Sense Induction, Web Search Result Clustering

1. INTRODUCTION

The use of word senses in place of surface word forms has been shown to improve performance on many computational tasks such as information extraction [5], information retrieval [20], data integration [3, 16], machine translation [21] and intelligent web search [14]. Two main techniques have been proposed to solve the word ambiguity problem from different perspectives: Word Sense Disambiguation (WSD) and Word Sense Induction (WSI). Word Sense Disambiguation is aimed at assigning word senses from a predefined *sense inventory* - such as the WordNet [13] database - to words in context while Word Sense Induction is based on clustering words according to their meanings without the use of a Sense Inventory. In other words, given a target word used in a number of different contexts, WSI is the process of clustering these instances of the target word together by determining which contexts are the most similar to each other.

In a previous paper [1] we experimentally evaluated whether the use of more than one language at a time for representing the context of a word may have positive effect on the performance of a Word Sense Induction task or, conversely, the noise increase invalidates the benefits given by a multilingual context representation. The experiments showed a clear overall improvement of the performance: the single-language setting is outperformed by the multi-language settings on almost all the considered words. The registered performance gain reaches peaks of about 40% with certain words and even if we consider the average F-Measure, the difference between the single-language case and the five-language case is about 5%.

In this paper we give some preliminary ideas to exploit our multilingual Word Sense Induction method to Web search results clustering. Web search result clustering aims to facilitate information search on the Web; rather than the results of a query being presented as a at list of text snippets, they are grouped on the basis of their similarity: each cluster is intended to represent a different meaning of the input query, thus taking into account the lexical ambiguity (i.e., polysemy) issue [12]. As observed in [12], many Web clustering engines group search results on the basis of their lexical similarity: as a result, text snippets with no word in common tend to be clustered separately even if they share the same meaning, whereas snippets with words in common may be grouped together even if they refer to different meanings of the input query.

The paper is organized as follows. In the rest of this Introduction a motivation example is given; in Section 2 the Multilingual Word Sense Induction (ML-WSI) method is outlined and, then, in Section 3 we discuss some preliminary ideas to exploit such ML-WSI method to Web search results clustering. Finally, in section 4, conclusions are drawn.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2743009.

1.1 Motivation Example

To give an intuition of the approach, let us consider the following keyword query: *protect snow leopard*; Google search returns, among others, the following snippets (subscript *EN* denotes that we are using the English language as target language):

$S1_{EN}$: World's leading authority on the study and protection of the endangered snow leopard.

$S2_{EN}$: One of the hidden features of Snow Leopard is a built-in system to protect Mac users from malware

$S3_{EN}$: The law especially protects Snow Leopards and hunting one is culpable of punishment by imprisonment and fines.

As observed in [12], although snippets 1 and 3 refer to the same meaning, they have no content word in common apart from our query words. As a result, a traditional Web clustering engine would most likely assign these snippets to different clusters. It has been shown in [12] that this problem can be addressed, thanks to a novel approach to Web search result clustering based on Word Sense Induction; the key of this approach is to first acquire the various senses (i.e., meanings) of an ambiguous query and then cluster the search results based on their semantic similarity to the word senses induced. The experiments, conducted on data sets of ambiguous queries, have shown that this approach outperforms both Web clustering and search engines.

The method we propose in [1] for WSI is based on the so-called *context clustering* approach [17]: the idea is that a given word - used in a specific sense - tends to co-occur with the same neighboring words [9]. In this approach each occurrence of a target word in a corpus is represented as a *vector of features*; the simplest features are the *unigrams* (individual words) or *bigrams* of words (couple of words) composing the context. In the example, we represent each snippet as a context vector, i.e., as a vector of words:

ctx_{1en} : { World, leading, authority, study, protection, endangered, snow leopard }

ctx_{2en} : { hidden, features, Snow Leopard, built-in, system, protect, Mac, users, malware }

ctx_3en : { law, especially, protects, Snow Leopards, hunting, culpable, punishment, imprisonment, fines }

These context vectors are clustered and the resulting clusters are taken to represent the induced senses. The performance of this clustering step highly depends on the quality of the features used for the context representation so the goal is to improve the quality of the features. In [1], we demonstrated that a *multilingual* context representation, where the context of the words is expanded by considering its translations in different languages, can improve the performance of the WSI process. In other words, the key of our approach is to perform Word Sense Induction on a *target language* by using *other languages* as support. To give an idea of our approach, by using the Italian language as support language, for the above three snippets we consider their Italian translation:

$S1_{IT}$: Delle principali autorita' del mondo sullo studio e la tutela del **leopardo delle nevi** in via di estinzione.

$S2_{IT}$: Una delle caratteristiche nascoste di **Snow Leopard** e' un sistema integrato per proteggere gli utenti Mac da malware.

$S3_{IT}$: La legge in particolare tutela **leopardi delle nevi** e caccia uno e' colpevole della pena con la reclusione e multe.

and their respective context vectors:

ctx_{1it} : { principali, autorita', mondo, studio, tutela, leopardo, nevi, via, estinzione }

ctx_{2it} : { caratteristiche, nascoste, Snow Leopard, sistema, integrato, proteggere, utenti, Mac, malware }

ctx_{3it} : { legge, particolare, tutela, leopardi, nevi, caccia, colpevole, pena, reclusione, multe }

As input of the clustering algorithm we then use the *multilingual context vector* obtained by the union of the english and italian context vectors. The use of this multilingual context representation may give a twofold improvement to the WSI task:

1. Enrichment of the number of features that can be used to distinguish contexts;

2. Improvement on the quality of the context representation.

Intuitively, it can be seen that *snow leopard* has been translated in different ways ('Snow Leopard', 'leopardo delle nevi') depending on the meaning that it assumes in the sentence; these translated words are included as features in the context vector representation and they may help the clustering process.

2. THE ML-WSI METHOD

In [1] we proposed the ML-WSI method which performs Word Sense Induction on words of a *target language* by using *other languages* as support; we applied such method by considering as target the English language and by using as support four languages, Italian, French, Spanish and Portuguese.

The ML-WSI method is based on the so-called *context clustering* approach composed by the following two steps [17]:

A) Context Representation
In the simplest representation, these context vectors are *first-order* vectors containing *unigrams* or *bigrams* of words, as previously shown in section 1.1. In a *second-order* context representation [8], used in our ML-WSI method, each word in a context to be clustered is replaced by a first-order vector; the vectors for all the words in a context are averaged together to create a single vector that becomes the second-order word by word co-occurrence representation of that context.

B) Context clustering
The next step of the Context clustering approach is to cluster the context representation obtained at the previous step; in this way, word's are grouped by meaning. We used a simple k-means algorithm since the idea proposed is independent from the algorithm selected.

The ML-WSI method is based on the construction of a *multilingual* second order context vector starting from a *parallel corpus*, i.e., a set of parallel aligned sentences in different languages. More precisely, a bilingual parallel corpus TL_SL, is organized as a set of *sentence pairs* $s_{i\_j} =< s_i, s_j >$, where s_i is a sentence in the target language TL and $s_j = T_j(s_i)$ is the corresponding translation in a support language SL. This definition is extended to an arbitrary number of support languages: $TL\_SL_1\_SL_2\_...\_SL_N = \{< s_{TL}, s_1, s_2, ..., s_N >| < s_{TL}, s_i > \in TL\_SL_i, 1 \leq i \leq N\}$. Since in our method more than two languages at a time need to be considered, we developed a tool able to generate *multilingual parallel* corpora, containing a virtually unlimited number of languages, starting from bilingual parallel corpora. In [1], for the experimental analysis, we used a multi-lingual parallel corpus JRC-Acquis [19]; this domain-specific corpus (it is a collection of European Union laws) is available in 22+ languages and the translation of the original text in all this languages has been conducted by expert translators and, thus, the results the translations are very accurate. In section 3.2 we will discuss the use of other Multilingual text corpus.

2.1 Experiments

To evaluate the effect of adding a language to the WSI task we considered four languages (IT-Italian, FR-French, ES-Spanish and PT-Portuguese) and we performed the following multi-lingual tests considering all the possible combinations of the selected languages:

- *Bi-lingual Tests*: all combinations of pairs of languages;
- *Tri-lingual Tests*: all combinations of triplets of languages;
- *Quadri-lingual Tests*: all combinations of quadruplets of languages;
- *Penta-lingual Test*: One final test with all five languages together.

The context representation models are built by considering for a word a sentence in English followed by the corresponding sentences in other languages.

In the context clustering step we followed standard unsupervised WSI settings :

- *testing set*: composed by the 5% of the corpus and manually annotated with the correct meaning, the so-called gold-standard sense;
- *training set*: constituted by the remaining 95% of the corpus and is obviously not annotated.

We have conducted test with 21 ambiguous target words, selected from the 30 words used in the SemEval 2010 competition by dropping those words with not enough instances within the JRC-Acquis corpus. The obtained results are graphically displayed in Figure 1 thus highlighting that by increasing the number of languages used, we obtained even better performance.

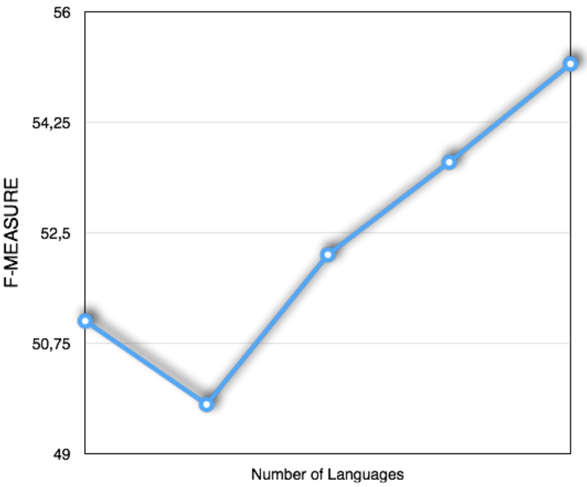

Figure 1: F-measure performance gain w.r.t. the number of languages

3. WEB SEARCH RESULTS CLUSTERING WITH ML-WSI

In this section we outline the main steps of the approach to Web search result clustering based on WSI proposed in [12]; then, we describe how such steps are modified to adopt our ML-WSI method. Our purpose is to extend the approach presented in [12] with our ML-WSI approach in order verify if the Web Search Clustering scenario can get benefits from the additional information introduced by a multilingual representation of the text.

Let Q_{EN} be a query in the target language (for sake of clarity we use the English language as target and the Italian language as support to the ML-WSI method).

1. all the possible word senses of Q_{EN} are induced by a WSI algorithm from a text corpus;

   With the ML-WSI algorithm, this step is performed by the following two steps:

   1a) Q_{EN} is translated in the support language by using a Machine Translation system (see section 3.1): Q_{IT};

   1b) all the possible word senses of $Q_{EN} \cup Q_{IT}$ are induced by the ML-WSI algorithm from a bilingual text corpus; (bilingual and multilingual text corpus are discussed in 3.2)

2. Q_{EN} is executed by the web search engine in order to get the search result snippets: $R = (S1_{EN}, S2_{EN}, ..., Sn_{EN})$; It should be noted that the query is only executed in the target language (English).

3. Each search snippet S_{EN} is processed and mapped to the most appropriate meaning by the WSI algorithm.

   In order to use the ML-WSI algorithm, this step requires that each snippet S_{EN} is translated in the support language by the MTS (**Snippet Translation** is discussed in Section 3.1).

4. The resulting clustering of snippets in

   $R = (S1_{EN}, S2_{EN}, ..., Sn_{EN})$ is returned.

In other words, the aim of our approach is to perform Web Search Results Clustering on English queries by using other languages as support: by submitting a query Q to a search engine, we obtain a list of relevant search results R.

The ML-WSI method we proposed in [1] relies on the translation process in order to obtain an increase in performance in the WSI task; a low quality translation may not only void the benefits given by the translation information but it may in addition lead to a loss of performance because of the noise introduced by wrongly translated words.

To exploit the ML-WSI method for Web search result clustering two fundamental ingredients are required:

1. a Machine Translation system, to translate in different languages both the query and the search snippets;
2. a multilingual and domain-independent corpus, because the web search clustering is a general-domain application and the use of a domain-specific corpus can adversely affect the clustering performance.

3.1 Query and Snippet Translation

To exploit the ML-WSI method for Web search result clustering a Machine Translation (MT) system is required to translate in different languages both the query and the search snippets. MT systems have worse performance compared to human translators; however, the performance of a state-of-the-art MT system should be enough for our purpose since we do not need a faultless translation. Moreover the fact that we translate the snippets and the query with the same MT system could be an advantage because the translations will probably be consistent. More specifically, translation is applied to snippets and not to the single words of a context features as phrase-based translation models outperform word-based translation models for almost all language pairs [10]. In our intuitive examples, we used one of the most common Machine Translation systems, the Google Translator (GT), which adopt a phrase-based translation model. It should be noted that our approach uses Machine Translation to improve the Word Sense Induction process. On the other hand, Machine Translation in an historical application for Word Sense Induction and a recent invited talk discussed the use of Word Sense Induction for Machine Translation [22].

Future works will be focused on the Snippet Translation process, a central step of the proposed method. First of all, we will evaluate the so-called Cross Lingual Snippet Generation systems [11], which generate snippets in multiple languages starting from documents available only in one language with the help of Machine Translation systems. Another interesting scenario to consider is when the same web page is available in more languages; in this case we will evaluate whether the Snippet Translation process may be performed by the so-called Multi Lingual Snippet Generation systems which are able to generate snippets in multiple languages [15].

3.2 Multilingual text corpus

The ML-WSI algorithm is based on the clustering of text snippets and for this reason it needs a large quantity of data in order to be trained. In a web search clustering scenario we are working with textual data, thus for the training of the algorithm we need a corpus, i. e. a large and structured set of texts; in particular, since we are working with different languages, we need a multilingual corpus. While there are many single language corpora (especially in english language), the multilingual corpora are few and their size is often not comparable with the typical size of a single language corpus.

Future works will follow two directions. First, we will consider general multilingual corpus, like DBPedia [2], and other multilingual corpus, like [18, 6]. Second, we will consider the automatic translation in different languages of the existent single language corpora, i.e., we will experimentally verify if the use of the multilingual corpus could be effectively replaced by the use of a MT system without a significative loss in performance for the specific application. In particular, to evaluate the impact of the ML-WSI method in the Web search result clustering, we will also consider the same two corpora used in [12]:

- **Google Web1T** [4]: This corpus is a large collection of n-grams ($n = 1, ..., 5$)-namely, windows of n consecutive tokens-occurring in one terabyte of Web documents as collected by Google.
- **ukWaC** [7]: This corpus was constructed by crawling the .uk domain and obtaining a large sample of Web pages that were automatically part-of-speech tagged using the TreeTagger tool. For this corpus we considered all the co-occurrences of WordNet lemmas that appear in the same sentence.

4. CONCLUSIONS

Word Sense Induction has been shown useful in many scenario. In [12], Word Sense Induction was proposed as a novel approach to Web search result clustering, in the context of an Intelligent Web Search scenario. In our previous paper [1] we proposed the ML-WSI method, an innovative Word Sense Induction method based on multilingual data. In the present paper, the adoption of the ML-WSI method to improve the performance of Web search result clustering is proposed and some issues for the future work are individuated.

We conclude with some considerations concerning the use of the ML-WSI method for Multilingual Web Access concerned with retrieval from the Web, where documents in multiple languages co-exist and need to be retrieved to a query in any language. It should be noted that, even if we talked about target and support language and our clustering considered only web search results in English, the method can be used also to retrieve documents in any language, because we translate both the query and the resulting snippets. This mean that we can execute one query for each considered language, and then we can translate the obtained snippets for that language, in all the other languages. In this way we obtain snippets written in the same set of languages, independently from the language used for the query and from the language of the retrieved document, and we can cluster by topic or meaning all the documents written in different languages.

5. REFERENCES

[1] L. Albano, D. Beneventano, and S. Bergamaschi. Word sense induction with multilingual features representation. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 2, pages 343–349. IEEE, 2014.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[3] D. Beneventano, S. Bergamaschi, and S. Sorrentino. Extending wordnet with compound nouns for semi-automatic annotation in data integration systems. In *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*, pages 1–8, Sept 2009.

[4] T. Brants and A. Franz. {Web 1T 5-gram Version 1}. *Linguistic Data Consortium, Philadelphia*, 2006.

[5] J. Chai and A. Biermann. The use of word sense disambiguation in an information extraction system. *Proceedings of Sixteenth National Conference in Artificial Intelligence and Eleventh Annual Conference on Innovative Applications of Artificial Intelligence*, July 1999.

[6] F. Eermak and A. Rosen. The case of intercorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427, 2012. cited By 1.

[7] A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, 2008.

[8] S. H. Automatic word sense discrimination. *Computational Linguistics*, 1998.

[9] Z. Harris. Distributional structure. 1954.

[10] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[11] P. Lohar, P. Bhaskar, S. Pal, and S. Bandyopadhyay. Cross lingual snippet generation using snippet translation system. In *Computational Linguistics and Intelligent Text Processing*, pages 331–342. Springer, 2014.

[12] A. D. Marco and R. Navigli. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, (July 2012), 2013.

[13] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[14] R. Navigli and M. Lapata. An experimental study on graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[15] S. B. Pinaki Bhaskar. Cross lingual query dependent snippet generation. *(IJCSIT) International Journal of Computer Science and Information Technologies*, 3 (4):4603–4609, 2014.

[16] L. Po and S. Sorrentino. Automatic generation of probabilistic relationships for improving schema matching. *Information Systems, Volume 36, Issue 2 (2011), pp. 192-208*, 2011.

[17] A. Purandare and T. Pedersen. Senseclusters - finding clusters that represent word senses. *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04)*, May 2004.

[18] R. Steinberger, M. Ebrahim, A. Poulis, M. Carrasco-Benitez, P. SchlÃijter, M. Przybyszewski, and S. Gilbro. An overview of the european unionâĂŹs highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707, 2014. cited By 0.

[19] R. Steinberger and et al. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy*, May 2006.

[20] O. Uzuner and B. Katz. Word sense disambiguation for information retrieval. *Proceedings of AAAI/IAAI1999*, July 1999.

[21] Vickrey, Biewald, Teyssier, and Koller. Word-sense disambiguation for machine translation. *HLT/EMNLP*, 2005.

[22] M. Zhang. Word sense induction for machine translation (invited talk). *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, December 2014.