Using Explicit Semantic Analysis for Cross-Lingual Link Discovery

Petr Knoth
KMi, The Open University
p.knoth@open.ac.uk

Lukas Zilka
KMi, The Open University
l.zilka@open.ac.uk

Zdenek Zdrahal
KMi, The Open University
z.zdrahal@open.ac.uk

Abstract

This paper explores how to automatically generate cross-language links between resources in large document collections. The paper presents new methods for Cross-Lingual Link Discovery (CLLD) based on Explicit Semantic Analysis (ESA). The methods are applicable to any multilingual document collection. In this report, we present their comparative study on the Wikipedia corpus and provide new insights into the evaluation of link discovery systems. In particular, we measure the agreement of human annotators in linking articles in different language versions of Wikipedia, and compare it to the results achieved by the presented methods.

1 Introduction

Cross-referencing documents is an essential part of organising textual information. However, keeping links in large, quickly growing, document collections up-to-date, is problematic due to the number of possible connections. In multilingual document collections, interlinking semantically related information in a timely manner becomes even more challenging. Suitable software tools that could facilitate the link discovery process by automatically analysing the multilingual content are currently lacking. In this paper, we present new methods for Cross-Lingual Link Discovery (CLLD) applicable across different types of multilingual textual collections.

Our methods are based on Explicit Semantic Analysis (ESA) introduced by Gabrilovich and Markovitch (2007). ESA is a method that calculates semantic relatedness of two texts by mapping their term vectors to a high dimensional space (typically, but not necessarily, the space of Wikipedia concepts) and by calculating the similarity between these vectors (instead of comparing them directly). The method has received much attention in the recent years and it has also been extended to a multilingual version called Cross-Lingual Explicit Semantic Analysis (CL-ESA) (Sorg and Cimiano, 2008). To the best of our knowledge, this method has not yet been applied in the context of automatic link discovery systems.

Since the CLLD field is relatively young, it is also important to establish a constructive means for evaluating these systems. Our paper provides insight into this problem by investigating the agreement/reliability of man-made links and by presenting a possible approach for the definition of ground truth, i.e. gold standard.

The paper brings the following contributions:

(a) It applies Explicit Semantic Analysis to the link discovery and CLLD tasks.

(b) It provides new insights into the evaluation of CLLD systems and into the way people link information in different languages, as measured by their agreement.

2 Related Work

CLLD Methods

Current approaches to link detection can be divided into three groups:

(1) *link-based* approaches discover new links by exploiting an existing link graph (Itakura and Clarke, 2008; Jenkinson et al., 2008; Lu et al., 2008).

(2) *semi-structured* approaches try to discover new links using semi-structured information, such as the anchor texts or document titles (Geva, 2007; Dopichaj et al., 2008; Granitzer et al., 2008; Milne and Witten, 2008; Mihalcea and Csomai, 2007).

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 2–10,
Chiang Mai, Thailand, November 8-12, 2011.

(3) *purely content-based* approaches use as an input plain text only. They typically discover related resources by calculating semantic similarity based on document vectors (Allan, 1997; Green, 1998; Zeng and Bloniarz, 2004; Zhang and Kamps, 2008; He, 2008). Some of the mentioned approaches, such as (Lu et al., 2008), combine multiple approaches. To the best of our knowledge, no approach has so far been reported to use Explicit Semantic Analysis to address this task.

The main disadvantage of the link-based and semi-structured approaches is probably the difficulty associated with porting them across different types of document collections. The two well-known solutions to monolingual link detection, the Geva's and Itakura's algorithms (Trotman et al., 2009), fit in these two categories. While these algorithms have been demonstrated to be effective on a specific Wikipedia set, their performance has significantly decreased when they were applied to a slightly different task of interlinking two encyclopedia collections. Purely content-based methods have been mostly found to produce slightly worse results than the two previous classes of methods, however their advantage is that their performance should remain stable across different document collections. As a result, they can always be used as part of any link discovery system and can even be combined with domain specific methods that make use of the link graph or semi-structured information. In practice, domain-specific link discovery systems can achieve high precision and recall. For example, *Wikify!* (Mihalcea and Csomai, 2007) and the link detector presented by Milne and Witten (2008) can be used to identify suitable anchors in text and enrich it with links to Wikipedia by combining multiple approaches with domain knowledge.

In this paper, we present four methods (three purely content-based and one combining the link-based and content-based approach) for CLLD based on CL-ESA. Measuring semantic similarity using ESA has been previously shown to produce better results than calculating it directly on document vectors using cosine and other similarity measures and it has also been found to outperform the results that can be obtained by measuring similarity on vectors produced by Latent Semantic Analysis (LSA) (Gabrilovich and Markovitch, 2007). Therefore, the cross-lingual extension of ESA seems a plausible choice.

Evaluation of link discovery systems

The evaluation of link discovery systems is currently problematic as there is no widely accepted gold standard. Manual development of such a standard would be costly, because: (a) the number of possible links is very high even for small collections, (b) the link generation task is subjective (Ellis et al., 1994) and (c) it is not entirely clear how the link generation task should be defined in terms of link granularity (for example, document-to-document links, anchor-to-document links, anchor-to-passage links etc.). Developing such a CLLD corpora manually would be even more complicated.

As a result, Wikipedia links were extracted and taken as the gold standard (ground truth) in a comparative evaluation in (Huang et al., 2008). The authors admit that Wikipedia links are not perfect (validity of existing links is sometimes questionable and useful links may be missing) the comparative evaluation of methods and systems should be considered informative only. For example, it would be naïve to expect that measuring *precision/recall* characteristics would be accurate.

In this paper we discuss the issues in automatically defining the ground truth for CLLD systems. We take into account the differences in the way people link content in different languages to assess the agreement between the different language versions with the goal to find out how well our system performs. Our experiments are conducted on the Wikipedia dataset, however we use the articles only as a set of documents abstracting from the Wikipedia encyclopedic nature.

3 The CLLD methods

This section describes the methods used in our experiments. The whole process of cross-language link detection is shown in Figure 1. The method takes as an input a new "orphan" document (i.e. a document that is not linked to other documents) written in the source language and automatically generates a ranked list of documents written in the target language (the suitable link targets from the source document). The task involves two steps: the *cross-language* step and the *link generation* step. We have experimented with four different CLLD methods: *CL-ESA2Links*, *CL-ESADirect*, *CL-ESA2ESA* and *CL-ESA2Similar* that will be described later on. The names of the methods

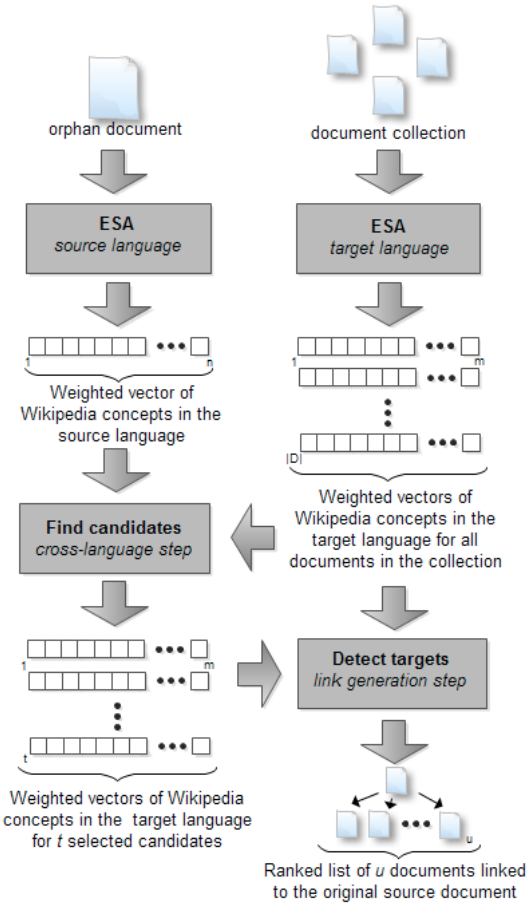

Figure 1: Cross-language link discovery process

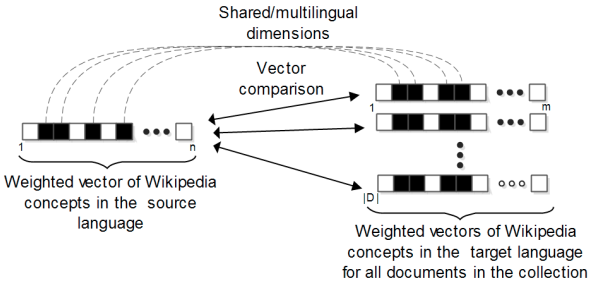

Figure 2: CLLD candidates

are derived from the approach applied in the first and the second step. These methods have different characteristics and would be useful in different scenarios.

In the **first step**, an ESA vector is calculated for each document in the document collection. This results in obtaining a weighted vector of Wikipedia concepts for each document in the target language. The cardinality of the vector is given by the number of concepts (pages) in the target language version of Wikipedia (i.e. it is about 3.8 million for English, 764,000 for Spanish, etc.). A similar procedure is applied on the orphan document, however, the source language version of ESA is used. The resulting ESA vector is then compared to the ESA vectors that represent documents in the target language collection (CL-ESA approach). A set of candidate vectors representing documents in the target language is acquired as an output of the cross-language step, see Section 3.1.

In the **second step**, the candidate vectors are taken as a seed and are used to discover documents that are suitable link targets. The four different approaches used in this step distinguish the above-mentioned methods, see Section 3.2.

3.1 The cross-language step

The main rationale for the cross-language step is to find t suitable candidates in the target language that can later be exploited to identify link targets. Semantically similar target language documents to the source language document are considered by our methods as suitable candidates. To identify such documents, the ESA vector of the source document is compared to the ESA vectors of documents in the target document collection.

Each dimension in an ESA vector expresses the similarity of a document to the given language version of a Wikipedia concept/article. Therefore, the cardinality of the source document vector is different from the cardinality of the vectors representing the documents in the target language collection (Figure 2). In order to calculate the similarity of two vectors, we map the dimensions that correspond to the same Wikipedia concepts in different language versions. In most cases, if a Wikipedia concept is mapped to another language version, there is a one-to-one correspondence between the articles in those two languages. However, there are cases when one page in the source language is mapped to more than one page in the target language and vice versa.[1] For the purpose of similarity calculation, we use 100 dimensions with the highest weight that are mappable from the source to the target language. The number of candidates to be extracted is controlled by parameter t. We have experimentally found that its selection has a significant impact on the performance of our methods.

[1]These multiple mappings appear quite rarely, e.g. in 5,889 cases out of 550,134 for Spanish to English and for 2,528 cases out of 163,715 for Czech to English.

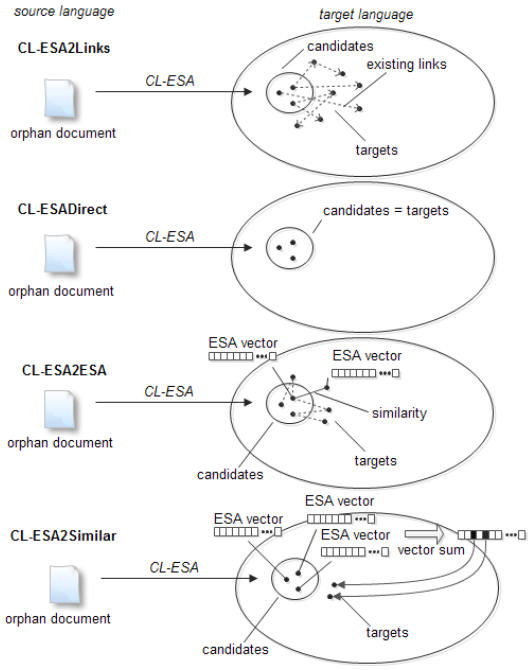

Figure 3: Schematic illustration of the four approaches used by the CLLD methods.

3.2 The link generation step

In the link generation step, the candidate documents are taken and used to produce a ranked list of targets for the original source document. The following approaches, schematically illustrated in Figure 3, are taken by our four methods:

- **CL-ESA2Links** - This method requires access to the link structure in the target collection. More precisely, the method takes the original orphan document in the source language and tries to link it to an already interlinked target language collection. After applying CL-ESA in the first step, existing links are extracted from the candidate documents. The link targets are then ranked according to their similarity to the source document, i.e. documents that are more similar are ranked higher. This list is then used as a collection of link targets.

- **CL-ESADirect** - This method applies CL-ESA on the source document and takes the list of candidates directly as link targets.

- **CL-ESA2ESA** - In this method, the application of CL-ESA is followed by another application of monolingual ESA, which measures the semantic similarity of the candidates with all documents in the document collection, to identify link targets.

- **CL-ESA2Similar** - Instead of generating the ranked list of link targets using monolingual ESA as in the previous method, which is computationally expensive, we calculate a vector sum from the candidate list of ESA document vectors. We then select strong Wiki concepts representing these dimensions as the set of targets. This is equivalent to calculating cosine similarity using *tfidf* vectors. Though much quicker, the main disadvantage is that if we wanted to use this method on another set than Wikipedia, ESA would have to be used with a different background collection.

All of the methods have different properties. CL-ESA2Links requires the knowledge of the link graph in the target document collection. CL-ESA2ESA and ESADirect are two methods that are universal, i.e. can be easily applied in any document collection. The difference between them is that the former one requires significantly less document vector comparisons than the later method. CL-ESA2Similar works almost as fast as CL-ESADirect, but it has the disadvantage that ESA has to be used with the specific document collection as a background.

4 The underlying data

Wikipedia has been used as a corpus for the methods evaluation. This decision has the following advantages that make it possible for us to test and analyse the methods on a real use case:

- A very large multilingual text collection.
- The articles are well-interlinked and the interlinking has been approved by a large community of users.
- A large proportion of articles contain explicit mapping between different language versions.

In our study, we have experimented with the English, Spanish and Czech language versions of Wikipedia. We consider the cases of linking from Spanish to English and from Czech to English, i.e. from a less resourced language to the more resourced one. We believe that this is the more interesting direction for CLLD methods as the target

language version is more likely to contain relevant information not available in the source language. The language selection has been motivated by the aim to test the methods in two very different environments. The Spanish version is relatively well resourced containing 764,095 pages (about four times fewer than English), the Czech language is much less resourced containing 196,494 pages (about four times fewer than Spanish).

5 Evaluation methodology

One of the main obstacles in systematically improving link discovery systems is the difficulty to evaluate the results. The issue that makes reliable evaluation problematic is due to both technical and cognitive aspects. The difficulty in obtaining the "ground truth" for a sufficiently large dataset is caused both by the lack of human resources to manually annotate a very large number of document combinations, and the inherent subjectivity of the task. As a result, we find it essential to estimate the agreement between annotators and see to what extent the precision and recall characteristics can be measured with respect to interlinked document collections.

We claim that the reasons for linking two pieces of information is made at the level of semantics, i.e. the annotator has to understand the concepts/ideas described in two papers to decide if they should be connected by a link. We claim that this process should be language independent. Thus, an article about London will be related to an article about the United Kingdom regardless of the language the articles are written in.

Therefore, let us define the link generation task in the following way: Given a document[2] in the source language, find documents in the target language that are suitable link targets for the source document, i.e. there is a semantic relationship between the source document and the linked target documents.

Based on the definition, the ground truth for a topic document d is the set of documents that can be considered (semantically) suitable link targets. Though this set is typically unknown to us, we can in our experiment approximate it by taking the existing Wikipedia links as ground truth. Because the Wikipedia link structure has been agreed by a large number of contributing authors, it is likely to have a relatively consistent link structure in comparison to content that would be linked just by a single person. To establish the ground truth for the original source document, we can extract all links originating in the source document and pointing to other documents. Since the process of linking information is performed at the semantic level, and is thus language independent, we can enrich our ground truth with link graphs from different language versions of Wikipedia. This causes the ground truth to get larger which has two consequences: (1) It increases the reliability of the evaluation as many relevant links are often omitted (Knoth et al., 2010) (2) It is more difficult to achieve higher recall.

[2]The term *topic* is also sometimes used to refer to the document.

6 Results

6.1 Experimental setup

The experiment was carried out for two language pairs: Spanish to English and Czech to English. We will denote the source language L_{source} and the target language L_{target}. The input for the different CLLD methods are two document sets:

- Let $SOURCE_{L_{source}}$ be the set of topic documents selected as pages that contain a Wikipedia link between different language versions. In our case, 100 pages were selected.
- Let $TARGET_{L_{target}}$ be the collection of documents in the target language from which the link targets are selected. In our case, this collection contains all (3.8 million) Wikipedia pages in English.

The output of the method is a set (ranked list) $LIST_{result} = \langle TARGET_{L_{target}}, score \rangle$. To establish the ground truth we define:

- Let ρ be the mapping from documents in the source language to their target language versions $\rho : D_{L_{source}} \rightarrow D_{L_{target}}$.
- Let $SOURCE_{L_{target}}$ be the set of topic documents mapped to the target language $SOURCE_{L_{target}} = \rho SOURCE_{L_{source}}$.
- Let α, β be the mappings from documents to the other documents they link to in the source and target language respectively $\alpha : D_{L_{source}} \rightarrow D_{L_{source}}$, $\beta : D_{L_{target}} \rightarrow D_{L_{target}}$.

then we define the ground truth (GT) as the union of ground truths for different language versions, in this experiment we define it as the union of ground truth for the source and target language.

$$GT = \alpha(SOURCE_{L_{source}}) \cup \beta(SOURCE_{L_{target}})$$

A given generated item $\langle d, score \rangle \in LIST_{result}$ is evaluated as a hit if and only if $d \in GT$.

6.2 Methods evaluation

To investigate the performance of the first part of CLLD - the cross-language step carried out by CL-ESA, we have analysed how well the system finds for a given topic document in the source language the duplicate document in the target language. In this step, the system takes a document in the source language, and selects from the 3.8 million large document set in the target language the documents with the highest similarity. We then check, if a duplicate document ($d = \rho d_{source}$) appears among the top k retrieved documents. The experiment is repeated for all examples in $SOURCE_{L_{source}}$ and the results are then averaged (Figure 4). The graph suggests that the method performs well, as the document often appears among the first few results. In about 65% of cases, the document is found among the first 50 retrieved items. We believe that if the set of candidates (controlled by the t parameter) contains this document, the CLLD method is likely to produce better results, this is especially true for the CL-ESA2Links method.

The overall results for all the methods are presented in Figure 5. We have experimentally set $t = 10$ for Spanish to English and $t = 3$ for Czech to English CLLD. CL-ESA2Links performed in the experiments the best achieving 0.2 precision at 0.3 recall. CL-ESA2Similar performed the best out of the purely content-based methods.

Though the precision/recall might seem quite low, a number of things should be taken into account:

- A significant number of potentially useful links is still missing in our ground truth, because people typically do not intend to link all relevant information. As a result, many potentially useful connections are not explicitly present in Wikipedia (Knoth et al., 2010).

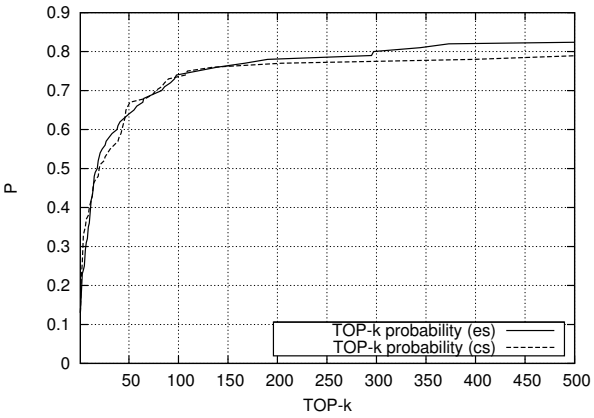

Figure 4: The probability (y-axis) of finding the target language version of a given source language document using CL-ESA in the top k retrieved documents (x-axis). Drawn as a cumulative distribution function.

  The problem can be partly mitigated by combining the ground truth from more language versions. Another approach is to measure the agreement instead of precision/recall characteristics (see Section 6.3).

- A significant number of links in Wikipedia are conceptual links. These links do not express a particularly strong relationship at the article level. This makes it very difficult for the pure-content based methods to find them, which results in low recall. It seems that CL-ESA2Links is the only method that does not suffer from this issue.

- The experiment settings make it hard for the methods to achieve high precision/recall performance. The $TARGET_{L_{target}}$ set contains 3.8 million articles, out of which, the methods are supposed to identify on average just a small subset of target documents. More precisely, in Spanish to English CLLD, our ground truth contains on average 341 target documents with standard deviation 293, in Czech to English, it contains on average 382 target documents with standard deviation 292.

6.3 Measuring the agreement

To assess the subjectivity of the link generation task and to investigate the reliability of the acquired ground truth, we have compared the link structures from different language version of

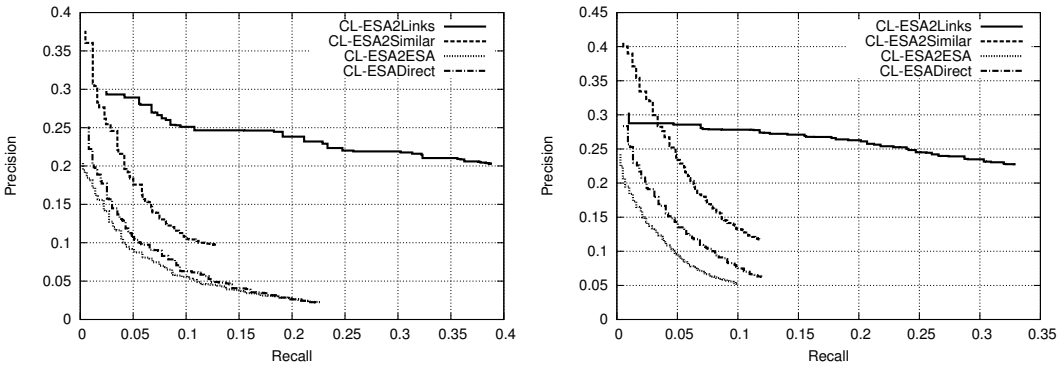

Figure 5: The precision ($y - axis$)/recall (x-axis) graphs for Spanish to English (left) and Czech to English (right) CLLD methods.

Spanish vs English			
	Y_{en}	N_{en}	N/A_{en}
Y_{es}	5,563	10,201	3,934
N_{es}	15,715	539,299,641	99,191,766
N/A_{es}	5781	321,326,145	0
Czech vs English			
	Y_{en}	N_{en}	N/A_{en}
Y_{cz}	4,308	8,738	2,194
N_{cz}	12,961	392,411,445	7,501,806
N/A_{cz}	9,790	356,532,740	0

Table 1: The agreement of Spanish and English Wikipedia and Czech and English Wikipedia on their link structures calculated and summed for all pages in $SOURCE_{es}$. Y - indicates yes, N - no, N/A - not available/no decision

Wikipedia. We have iterated over the set of topics from $SOURCE_{L_{source}}$ and recorded for each document in $TARGET_{L_{target}}$ in each step if it is a valid link target (yes - Y) or if it is not a valid link target (no - N) for the given source document in each language, thus measuring the agreement between the link structures in different languages. The results are presented in Table 1.

As demonstrated in Figure 6, a subset of Wikipedia pages cannot be mapped to other language versions. Either the semantically equivalent page does not exist or the cross-language link is missing. These links were classified as no decision/not available (N/A). The mappable documents were classified in a standard way according to their appearance in the link graphs of the language versions. Only these links are taken into account while measuring the agreement.

A common way to assess inter-annotator agreement between two raters in Information Retrieval is using the Cohen's Kappa calculated as:

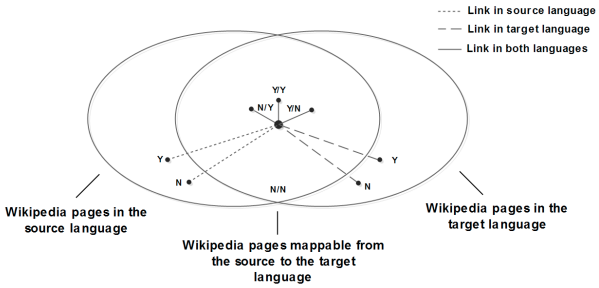

Figure 6: Individual cases of agreement/disagreement/no decision (not available) for two language versions of Wikipedia link graphs.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

where $Pr(a)$ is the relative observed frequency of agreement and $Pr(e)$ is the hypothetical probability of chance agreement. $Pr(a)$ is typically calculated as $\frac{|Y,Y|+|N,N|}{|Y,Y|+|Y,N|+|N,Y|+|N,N|}$. Since there is a strong agreement on the negative decisions, the probability will be close to 1. If we ignore the $|N,N|$ cases, which do not carry any useful information, the formula looks as follows:

$$Pr(a) = \frac{|Y,Y|}{|Y,Y| + |Y,N| + |N,Y|}.$$

The probability of a random agreement is extremely low, because the probability of a link

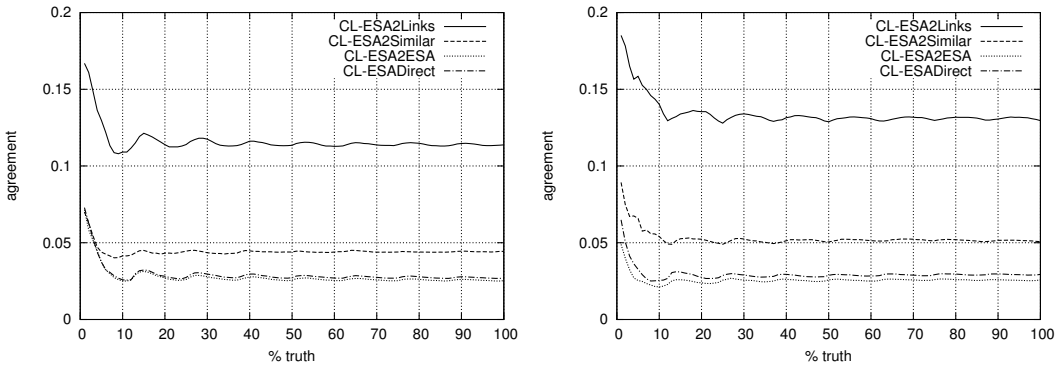

Figure 7: The agreements of the Spanish to English (left) and Czech to English (right) CLLD methods with $GT_{es,en}$ and $GT_{cz,en}$ respectively. The y-axis shows the agreement strength and the x-axis the number of generated examples as a fraction of the number of examples in ground truth.

connecting any two pages is approximately:[3]

$$p_{link} = \frac{|links|}{|pages|^2} = \frac{78.3M}{3.2M^2} = 0.000007648.$$

Thus, the hypothetical number of items appearing in the Y, Y class by chance is $p_{link}^2 \cdot (|Y, Y| + |Y, N| + |N, Y| + |N, N|)$. This formula estimates the number of agreements achieved by chance. In our case the value is much smaller than 1, hence $P(e)$ is close to 0. Therefore, we can calculate the agreement for English and Spanish as:

$$\kappa_{en,es} = \frac{5,563}{31,479} = 0.177.$$

The agreement for Czech and English is:

$$\kappa_{en,cz} = \frac{4,308}{26,007} = 0.166.$$

The value indicates a relatively low inter-annotator agreement. We believe that the fact that such a low agreement has been measured is very interesting, particularly because the link structure in Wikipedia is a result of a collaborative effort of many contributors. Therefore, we would expect that even lower agreement might be experienced in other types of text collections.

Motivated by the previous findings, we have calculated the agreement between the output of our method and the link graphs present in different language versions of Wikipedia. We were especially interested to find out if the agreement is significantly different from the agreement measured between different language versions of Wikipedia. We have generated by our CLLD methods 100% of $|GT|$ links for every orphan document in $SOURCE_{L_{source}}$, i.e. if a particular document is linked in Wikipedia to 57 documents, we generate 57 links. We have then measured the agreement for each topic document and averaged the agreement values. The results of the experiment for Spanish to English and Czech to English CLLD are shown in Figure 7. They suggest that CL-ESA2Links achieved a level of agreement comparable to that of human annotators. A very reasonable level of agreement has also been measured for CL-ESA2Similar, especially for the first 10% of the generated links. CL-ESADirect and CL-ESA2ESA exhibit a lower level of agreement.

7 Conclusion

In this paper, we have presented and evaluated four different methods for Cross-Language Link Discovery (CLLD). We have used Cross-language Explicit Semantic Analysis as a key component in the development of the four presented methods. The results suggest that methods that are aware of the link graph in the target language achieve slightly better results than those that identify links in the target language only by calculating semantic similarity. However, the former methods cannot be applied in all document collections and thus the later methods are valuable. Though it might seem at first sight that CLLD methods do not provide very high precision and recall, we have shown that the performance can, in fact, reach the results achieved by human annotators.

[3] Following the official Wikipedia statistics. Though different language versions have different p_{link}, the differences do not effect the results.

References

James Allan. 1997. Building hypertext using information retrieval. *Inf. Process. Manage.*, 33:145–159, March.

Philipp Dopichaj, Andre Skusa, and Andreas Heß. 2008. Stealing anchors to link the wiki. In Geva et al. (Geva et al., 2009), pages 343–353.

David Ellis, Jonathan Furner-Hines, and Peter Willett. 1994. On the measurement of inter-linker consistency and retrieval effectiveness in hypertext databases. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 51–60, New York, NY, USA. Springer-Verlag New York, Inc.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.

Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors. 2009. *Advances in Focused Retrieval, 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008, Dagstuhl Castle, Germany, December 15-18, 2008. Revised and Selected Papers*, Lecture Notes in Computer Science. Springer.

Shlomo Geva. 2007. Gpx: Ad-hoc queries and automated link discovery in the wikipedia. In Norbert Fuhr, Jaap Kamps, Mounia Lalmas, and Andrew Trotman, editors, *INEX*, Lecture Notes in Computer Science. Springer.

Michael Granitzer, Christin Seifert, and Mario Zechner. 2008. Context based wikipedia linking. In Geva et al. (Geva et al., 2009), pages 354–365.

Stephen J. Green. 1998. Automated link generation: can we do better than term repetition? *Comput. Netw. ISDN Syst.*, 30(1-7):75–84.

Jiyin He. 2008. Link detection with wikipedia. In Geva et al. (Geva et al., 2009), pages 366–373.

Wei Che Huang, Andrew Trotman, and Shlomo Geva. 2008. Experiments and evaluation of link discovery in the wikipedia.

Kelly Y. Itakura and Charles L. A. Clarke. 2008. University of waterloo at inex 2008: Adhoc, book, and link-the-wiki tracks. In Geva et al. (Geva et al., 2009), pages 132–139.

Dylan Jenkinson, Kai-Cheung Leung, and Andrew Trotman. 2008. Wikisearching and wikilinking. In Geva et al. (Geva et al., 2009), pages 374–388.

Petr Knoth, Jakub Novotny, and Zdenek Zdrahal. 2010. Automatic generation of inter-passage links based on semantic similarity. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 590–598, Beijing, China, August.

Wei Lu, Dan Liu, and Zhenzhen Fu. 2008. Csir at inex 2008 link-the-wiki track. In Geva et al. (Geva et al., 2009), pages 389–394.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA. ACM.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, *CIKM*, pages 509–518. ACM.

Philipp Sorg and Philipp Cimiano. 2008. Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes for the CLEF 2008 Workshop*.

Andrew Trotman, David Alexander, and Shlomo Geva. 2009. Overview of the inex 2010 link the wiki track.

Jihong Zeng and Peter A. Bloniarz. 2004. From keywords to links: an automatic approach. *Information Technology: Coding and Computing, International Conference on*, 1:283.

Junte Zhang and Jaap Kamps. 2008. A content-based link detection approach using the vector space model. In Geva et al. (Geva et al., 2009), pages 395–400.