A Comparative Evaluation of Cross-lingual Text Annotation Techniques

Lei Zhang[1], Achim Rettinger[1], Michael Färber[1], and Marko Tadić[2]

[1] Institute AIFB, Karlsruhe Institute of Technology, Germany
[2] Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
{l.zhang,rettinger,michael.faerber}@kit.edu,
{marko.tadic}@ffzg.hr

Abstract. In this paper, we study the problem of extracting knowledge from textual documents written in different languages by annotating the text on the basis of a cross-lingual knowledge base, namely Wikipedia. Our contribution is twofold. First, we propose a novel framework for evaluating cross-lingual text annotation techniques, based on annotation of a parallel corpus to a hub-language in a cross-lingual knowledge base. Second, we investigate the performance of different cross-lingual text annotation techniques according to our proposed evaluation framework. We perform experiments for an empirical comparison of three approaches: (i) *Cross-lingual Named Entity Annotation* (CL-NEA), (ii) *Cross-lingual Wikifier Annotation* (CL-WIFI), and (iii) *Cross-lingual Explicit Semantic Analysis* (CL-ESA). Besides establishing an evaluation framework, our results show the advantages and disadvantages of the three investigated approaches and clarify the roles of them for different purposes.

1 Introduction

Text annotation is about attaching additional information such as attributes, comments, descriptions, tags or links to a document or to textual units like words and phrases. In contrast to linguistic processing of natural language text, such as part-of-speech (POS) tagging and named entity recognition and classification (NERC), text annotation studied in this paper goes one level deeper. It enriches unstructured text with links to a knowledge base. In this regard, text annotation helps to bridge the gap between the ambiguity of natural language text and the corresponding formal representations in knowledge bases.

Text annotation as it is understood in this paper is defined in two ways: (i) linking entity mentions in documents to their corresponding representations in the knowledge base; (ii) linking the documents by topics to the relevant resources in the knowledge base. *Cross-lingual* text annotation becoming more and more popular goes beyond general annotation, as it faces the task of linking entities and topics across the boundaries of languages. Here, the text to be annotated and the resources in the knowledge base might be of different languages. In order to manage this new situation, a central knowledge base, where all entities are ultimately linked to, is needed. In our case, Wikipedia was chosen, as it is the largest on-line encyclopaedia up to date. Its articles are contributed by millions of users over the Web and cover any entity or topic of interest for most end users over the world. In addition, Wikipedia articles that provide information about the same concept in different languages are connected through cross-language links. A wide range of applications can benefit from its multilingualism.

Within the context of globalization, mainly driven by the digital revolution, institutions of any kind can no longer focus only on documents written in one language, but instead operate in various markets in different languages. In such a globalized and multilingual society, cross-lingual text annotation is crucial for processing natural language text in many different tasks. The following scenarios illustrate its application potentials:

- *Entity Tracking*: A business news website provides current statistics about companies around the world. For each company a dedicated web page displays a list of up-to-date relevant news articles that mention the company. It is essential to detect mentions of each company in the real-time multilingual news streams and to provide the latest relevant company news, preferably from their home markets. This is the task called *entity tracking*.
- *Topic Detection*: For a press agency, it is extremely important to determine the topic coverage of its news articles. As such, detecting the current topics from the global news streams, especially in different languages, is a task of great significance called *topic detection*. It can provide the editors with better understanding of recent developments in the global news topics and will indicate demand on the publishing market – i.e., what the publisher should write about because it is relevant to their audience and not yet or poorly covered from a global perspective.
- *Cross-lingual Recommendation*: An on-line news delivery service recommends relevant articles to its users around the world using materials previously read by the users as the context. To cater for its global customer readership, this service processes the multilingual news streams and provides *cross-lingual recommendations*, the task of finding relevant articles in different languages.

These scenarios described above motivate our study of cross-lingual text annotation in this paper. Regarding the *entity tracking* scenario, due to the general applicability of Wikipedia which contains an enormous number of entities in diverse domains, there is no problem to define the interests of the customers as a set of Wikipedia pages[3]. As a consequence, statements about whether specific newswire articles written in different languages are of interest can be made by linking entity mentions to the corresponding Wikipedia pages. In addition, Wikipedia covers a wide range of topics[4]. Therefore, cross-lingual text annotation can also be employed for *topic detection* by linking articles to their Wikipedia topics. In the case of *cross-lingual recommendation*, a measure to compute the

[3] E.g. `http://en.wikipedia.org/wiki/Deutsche_Bank` represents Deutsche Bank AG, the German global banking and financial services company.

[4] Topics such as, but not limited to, arts, history, events, geography, mathematics, and technology.

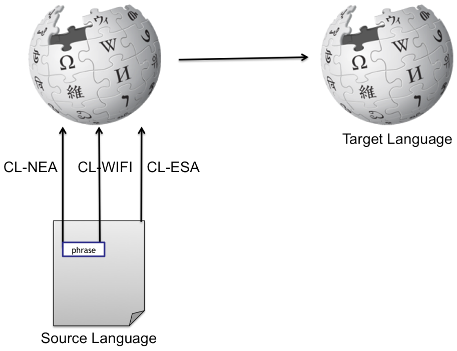


Fig. 1: Approaches for Cross-lingual Text Annotation.

similarity of texts in different languages is needed. However, due to the vocabulary mismatch problem, we cannot compare them directly. Through the annotation with Wikipedia, the documents in different languages will be first mapped to the entities or topics in a hub language in the knowledge base, e.g. English Wikipedia, before they can be compared.

The remainder of the paper is structured as follows: In Section 2, we present the approaches for cross-lingual text annotation studied in this paper. In Section 3, we describe our data, evaluation setting, and results followed by conclusions in Section 4.

2 Techniques for Cross-lingual Text Annotation

In this section, we present three approaches: (i) *Cross-lingual Named Entity Annotation* (CL-NEA) based on named entity recognition and classification (NERC) techniques, (ii) *Cross-lingual Wikifier Annotation* (CL-WIFI) based on the state-of-the-art wikification system, and (iii) *Cross-lingual Explicit Semantic Analysis* (CL-ESA) based on the Explicit Semantic Analysis (ESA) method. It should be noted that for CL-NEA the NERC systems are trained for each language individually on the annotated data. In contrast, CL-WIFI and CL-ESA are directly trained on Wikipedia. Fig. 1 illustrates these three approaches mentioned above. It is observed that all of them make use of the cross-language links in Wikipedia to find the corresponding Wikipedia pages in the different target languages. In the following, we briefly describe these approaches.

2.1 Cross-lingual Named Entity Annotation

Named entity recognition and classification (NERC) is the task within the field of information extraction (IE) of detecting specific information units within text such as names of persons, organizations, and locations. Since its beginnings in the early 1990s, NERC tools primarily have focused on these few classes: Per, Loc, Org, and Misc. During this time span, the focus evolved from rule-based algorithms to more and more machine learning techniques. In the following,

Table 1: Excerpt of the CoNLL 2003 data set. The first item on each line is a word, the second the corresponding part-of-speech (POS) tag, the third a syntactic chunk tag and the fourth the named entity tag.

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

we confine ourselves to supervised machine learning NERC techniques. They can be differentiated by the underlying model they use: Hidden Markov Model (HMM) [1], Decision Tree [2], Maximum Entropy Model (MEM) [3], Support Vector Machine (SVM) [4], or Conditional Random Field (CRF) [5].

For all supervised learning methods, appropriate training data is needed. Table 1 gives an impression of how such a training corpus for NERC can look like. For each term in a sentence, annotation in the form of a POS tag, a syntactic tag, and a NE tag has to be provided.

In our case, NERC for English and Spanish is performed by using AdaBoost on decision trees as described by Carreras et al. [6]. Carreras' approach has obtained best results in the CoNLL-2002 named entity extraction task and treats named entity recognition (NER) and named entity classification (NEC) as two separate tasks which are processed sequentially and independently. NER is performed as a combination of three local classifiers. These classifiers test simple hold decisions on each word in the text. For each target word several features such as lexical, syntactic, orthographical, and affix features are used. The task of NEC is to assign an entity type to an already found named entity and the multiclass multilabel AdaBoost.MH algorithm [7] is used. NEC is modeled here as a four-class classification problem with the four classes Per, Org, Loc, and Misc. Training was performed by using the CoNLL 2003 data set[5] for English and an updated version of the CoNLL 2002 shared task data set for Spanish (today included in the corpus Ancora[6]). NERC for the German language is performed by using the Stanford NERC tool which is based on the conditional random field model. For training, the CoNLL 2003 data set was used again. For more information, see [8].

On top of the standard monolingual NERC processing, a straight-forward approach for finding the corresponding Wikipedia page in another language is deployed: at first, the NE string is used for a keyword search for the Wikipedia article in the same language having the NE as title; then the cross-language links of this Wikipedia page are used to find the corresponding Wikipedia article of the target language (here, English). NERC is used here as computationally

[5] http://www.cnts.ua.ac.be/conll2003/ner/

[6] http://clic.ub.edu/ancora

inexpensive, but viable way for entity recognition and classification and as a prerequisite for cross-lingual entity linking.

2.2 Cross-lingual Wikifier Annotation

The process of augmenting phrases in text with links to their corresponding Wikipedia articles (in the sense of Wikipedia article annotation) is known as *wikification.* Training can here be performed on a Wikipedia dump directly. This means that we do not need any gold standard for the annotation of POS, syntactic chunk or NE tags for training, but only Wikipedia as corpus.

While Mihalcea and Csomai [9] met the challenge of wikification by using link probabilities obtained from Wikipedia's articles and by a comparison of features extracted from the context of the phrases, Milne and Witten [10] could improve the wikification service significantly by viewing wikification even more as a supervised machine learning task: Wikipedia is used here not only as a source of information to point to, but also as training data used to find always the appropriate link. Due to the richness of intra-wiki links and the large size of the English Wikipedia, evaluation showed better performance.

Entity linking in general consists of two main steps: entity detection and disambiguation. While disambiguation ensures that the detected phrases link to the correct entity (here: Wikipedia article) and therefore normally has to be done after entity detection, Milne and Witten let the disambiguation training phase be a prerequisite for detection.

Regarding training for disambiguation, three features are used: commonness, relatedness, and goodness of the context. The commonness of a candidate phrase is representing the proportion of linkage to the corresponding Wikipedia page in comparison to other link targets. With the help of the relatedness feature, the semantic context of the candidate phrase is taken into consideration. The relatedness is measured by the Google similarity distance (GSD) [11]. Since not all context terms are equal, but instead some are more meaningful, each context term is given a specific weight. By summing up the weights of the context terms, a feature context quality representing the goodness of the context can be generated. Based on these features, a classifier can be trained for disambiguation. The machine learning based link detection makes use of several features: link probability, relatedness, disambiguation confidence, generality, and location and spread. In this way, the context terms are used for learning what terms should and what should not be linked to.

As already presented before, linkage into another language is done by the cross-language links in Wikipedia.

2.3 Cross-lingual Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) has been proposed recently as an alternative approach for semantic modeling of natural language by exploiting unstructured or semi-structured text corpora instead of the traditional hand-crafted resources such as WordNet, taxonomies, or ontologies. Based on a given set of concepts

Table 2: Statistics about Wikipedia

(a) Number of articles.

	English Wikipedia	German Wikipedia	Spanish Wikipedia
#Articles	4,014,643	1,438,325	896,691

(b) Number of cross-language links.

	English-German	English-Spanish	German-Spanish
#Links (→)	721,878	568,210	295,415
#Links (←)	718,401	581,978	302,502
#Links (merged)	722,069	593,571	307,130

with textual descriptions, ESA defines the representation of documents with respect to these concepts. Various knowledge sources for concept definitions have been used. One of the most prominent examples is Wikipedia [12,13]. Concepts are hereby defined by Wikipedia articles, each of which comprises a comprehensive exposition of a topic.

ESA has been successfully applied to compute semantic relatedness between texts [12] or in text categorization tasks [13]. In the context of the cross-language information retrieval (CLIR) task, ESA has been extended to a cross-lingual setting (CL-ESA) by mapping the semantic document representation from one Wikipedia space to a Wikipedia space of another language [14,15]. This is achieved by exploiting language links in Wikipedia. As we use this approach as our third one for cross-lingual annotation, we briefly describe the underlying theory in the following:

Essentially, CL-ESA takes as input a document $d_s \in D_s$ in the source language L_s and maps it to a high-dimensional real-valued vector space spanned by a Wikipedia database $W_t = \{a_1, \ldots, a_n\}$ in the target language L_t such that each dimension corresponds to an article a_i acting as a concept. In this sense, the semantic representation of document d_s defined by concepts in W_t is given by the mapping function

$$\Phi(d_s) = [\phi(\tau_{t \to s}(a_1), d_s), \ldots, \phi(\tau_{t \to s}(a_n), d_s)]^T$$

where $\tau_{t \to s}(a_i)$ maps the Wikipedia article a_i in language L_t to the corresponding article in Wikipedia database W_s for language L_s. $\phi(a, d)$ denotes the strength of association between the document d and the Wikipedia article a in the same language, which can be defined using a tf-idf function based on the bag-of-words model [14]. Due to the large number of Wikipedia articles, in practice we consider only the top-k dimensions of the vector yielded by CL-ESA with the highest values. In our experiments, we set $k = 100$.

3 Experimental Evaluation

In this section, we propose a novel framework for evaluating cross-lingual text annotation techniques. According to this framework, we perform experiments to

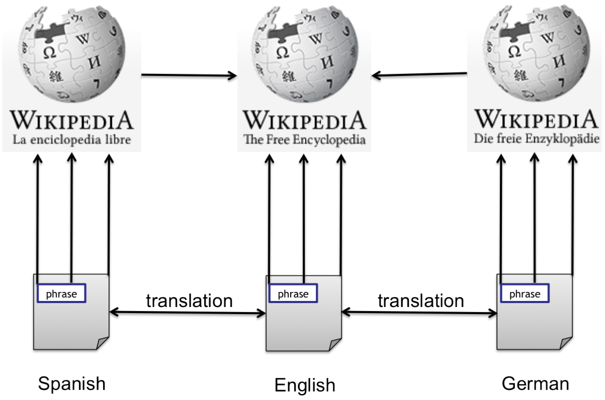


Fig. 2: Evaluation setting.

investigate the performance of different approaches. The goal of the evaluation is to measure the cross-lingual linking capabilities of the discussed approaches w.r.t. the annotations (links) of test documents in the source language (here English, German and Spanish) to the corresponding Wikipedia articles in the target language (English).

In the following, we first introduce the evaluation setting. Then, we provide the evaluation results of the three approaches (CL-NEA, CL-WIFI and CL-ESA). Our focus is on an empirical comparison of these approaches.

3.1 Evaluation Setting

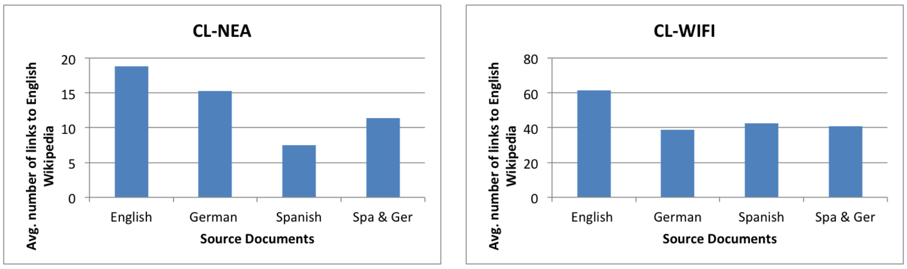
For the purpose of evaluation, we make use of a random sample of documents in English, German and Spanish from a parallel corpus[7] as test collection. While the evaluation of CL-NEA and CL-WIFI is focused on annotating word phrases in the test documents and linking each phrase to a single Wikipedia article describing it, CL-ESA is evaluated by linking each test document to a certain number of Wikipedia articles which are topically relevant. The evaluation setting is illustrated in Fig. 2.

To provide the test documents, we use the parallel corpus JRC-Acquis[8], which consists of legislative documents from the European Union and is widely used in cross-lingual research fields. The corpus is available in 22 European languages and comprises of approximately 23,000 documents in each language. In our experiments, we randomly select 88 parallel English-German-Spanish documents, each of which contains the translations of the same document in the above three languages.

Wikipedia is currently the largest knowledge base on the web and various editors develop it constantly, therefore its breadth and depth are expanding continually. The Wikipedia articles are available in approximately 270 languages

[7] Parallel corpus contains translated equivalents of documents in different languages.

[8] http://langtech.jrc.it/JRC-Acquis.html



(a) Number of CL-NEA links.

(b) Number of CL-WIFI links.

Fig. 3: Number of links detected by different approaches.

and they are linked to each other via cross-language links in case they describe the same topic. Most Wikipedia articles are available in English (currently more than 4 million pages). The advantage of Wikipedia is that the articles are not only available in a vast amount with regard to the number of pages per language, but also with regard to the number of different domains in its different languages. That is why we use Wikipedia as our nucleus[9].

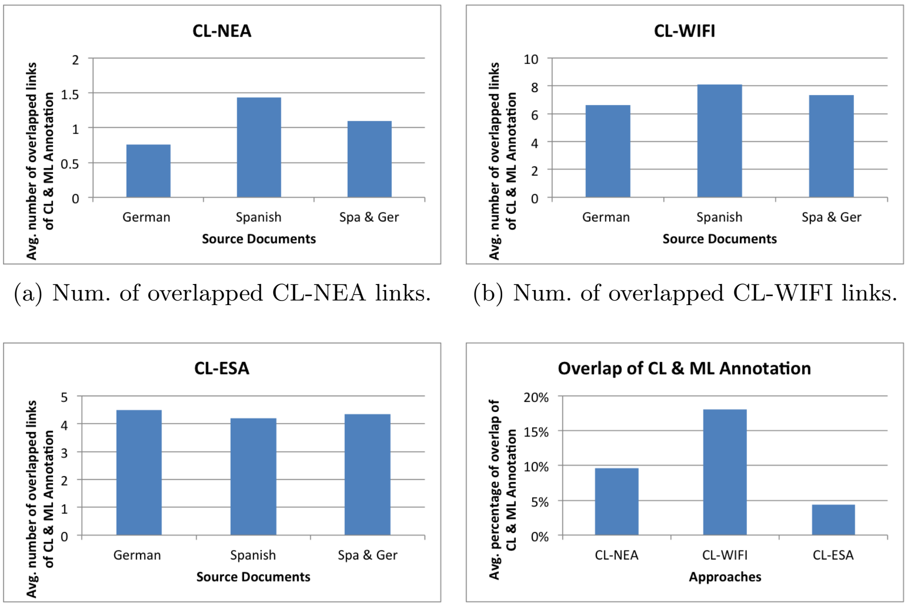
Table 2 shows some statistics of the Wikipedia articles in English, German and Spanish as well as the cross-language links between the articles in these languages extracted from Wikipedia snapshots of May 2012, which are used in our experiments. We analyze cross-language links between Wikipedia articles for each pair of supported languages in both directions and keep only articles for which aligned versions exist at least in one direction. For instance, we have extracted 721,878 cross-language links from English to German, and 718,401 links from German to English. By merging them together, we obtain 722,069 cross-language links, which are used to construct the cross-lingual knowledge base of the English-German language pair.

3.2 Evaluation Results

At first, we count the number of links to the English Wikipedia detected by each approach. Fig. 3a shows the average number of links per document detected by CL-NEA for different source languages. The results of CL-WIFI are shown in Fig. 3b. Concerning CL-ESA, we study whether the top-100 linked English Wikipedia topics are relevant to each test document. Therefore, the average number of detected links for each source language is 100.

It is expected that monolingual annotation of English documents detects more links than cross-lingual annotation of German/Spanish documents. This is due to the imbalance in the contents of Wikipedia in different languages and the missing cross-language links. In other words, English Wikipedia contains more articles, and not all Wikipedia articles in other languages are connected with their corresponding English versions. As shown in Fig. 3, for both CL-NEA and CL-WIFI, more links are detected in English documents by monolingual

[9] The Wikipedia database dumps are available at `http://dumps.wikimedia.org/`.



(a) Num. of overlapped CL-NEA links.

(b) Num. of overlapped CL-WIFI links.

(c) Num. of overlapped CL-ESA links.

(d) Percentage of overlapped links.

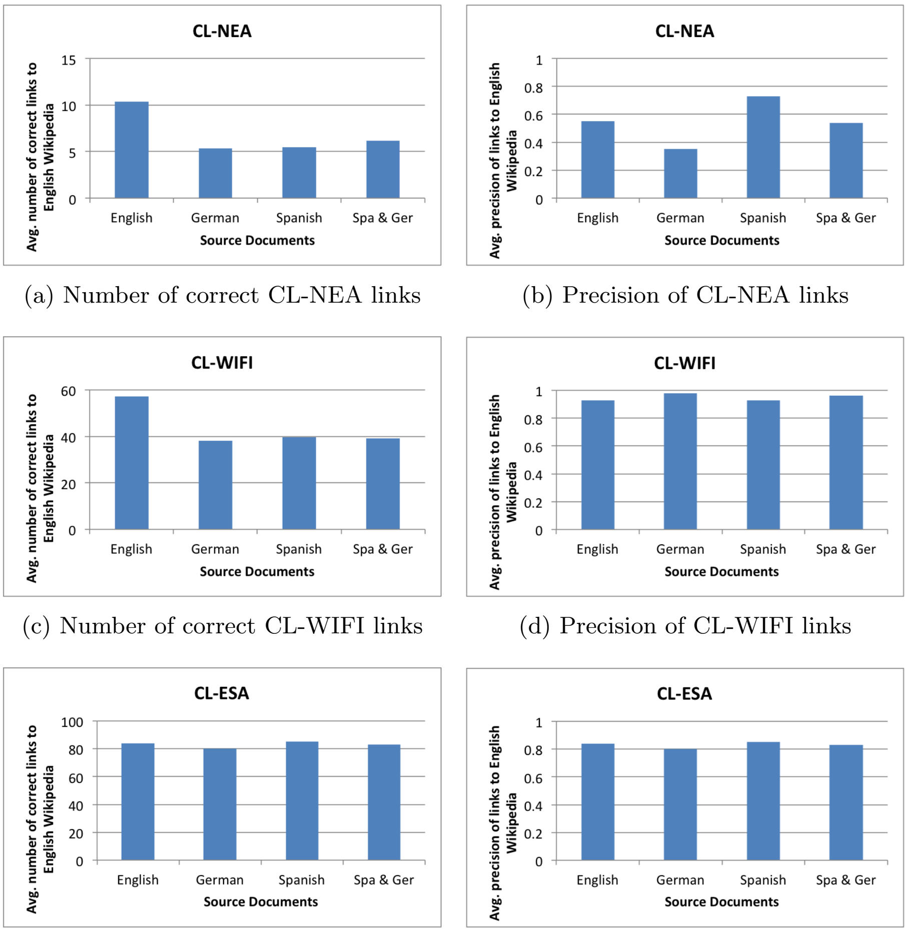
Fig. 4: The gap between cross-lingual and monolingual annotation.

annotation compared to cross-lingual annotation of German/Spanish documents, which conforms to our expectation.

It should be noted that CL-WIFI produces many more annotations than CL-NEA. The reason for that as we believe is that CL-WIFI is trained directly on Wikipedia, while CL-NEA is firstly trained on some other data sets before the detected entities are grounded in Wikipedia in a second step. In this sense, a lot of entities covered in Wikipedia might be missing in the training data sets used by CL-NEA.

Further, we try an automatic processing by comparing the links to English Wikipedia detected by cross-lingual annotation of German/Spanish documents with the ones found by monolingual annotation of English documents. Since this processing was done on a collection of parallel documents, it is expected that the same annotations should be found in any language, which makes the detected links comparable.

However, the number of the same links found by both cross-lingual and monolingual annotation indicates a low overlap between them. Fig. 4a shows the average number of overlapped CL-NEA links detected in both German/Spanish and English documents. The results of CL-WIFI and CL-ESA are illustrated in Fig. 4b and Fig. 4c, respectively. As shown in Fig. 4d, the average percentages of the overlapped links based on CL-NEA, CL-WIFI and CL-ESA are 9.6%, 18.1% and 4.4%, respectively. In general, we believe that the content imbalance and the missing cross-language links in Wikipedia used by cross-lingual annotation is also the reason of such a low overlap for all approaches. Compared with CL-WIFI,



(a) Number of correct CL-NEA links

(b) Precision of CL-NEA links

(c) Number of correct CL-WIFI links

(d) Precision of CL-WIFI links

(e) Number of correct CL-ESA links

(f) Precision of CL-ESA links

Fig. 5: Performance of different approaches

the percentage achieved by CL-NEA is much lower. That is because CL-NEA is trained on the data sets that contain completely different named entities for each language while CL-WIFI is trained directly on Wikipedia in which there exists a larger overlap among the articles in different languages. It might seem less intuitive that CL-ESA which is also trained on Wikipedia even yields a lower percentage than CL-NEA. This is due to the fact that CL-ESA links the documents to the Wikipedia articles by topics based on the bag-of-words model. In such a coarse-grained manner, the specific contextual words in different languages increase the gap between cross-lingual and monolingual annotation in an unexpected way.

In addition to the automatic evaluation, we also investigate the performance of different approaches by a manual evaluation w.r.t. the number of correct links and the precision of detected links, i.e. the fraction of the correct ones. In this regard, the detected links to the English Wikipedia for each source language were manually evaluated by marking the correctness of them.

Figs. (5a+5c+5e) illustrate the number of correct links detected by each approach. Clearly, CL-ESA produces more correct links than CL-WIFI, which in turn finds more correct ones than CL-NEA. The average precision of links detected by CL-NEA is shown in Fig. 5b. The results of both cross-lingual and monolingual annotation are somewhat below our expectation. We believe the reason of less correct links and lower precision yielded by CL-NEA in comparison to the other approaches is still the distinction between its training data and Wikipedia. In contrast, the average precision obtained by CL-WIFI, as shown in Fig. 5d, exceeds 0.9 for all three languages. Fig. 5f shows the precision of CL-ESA links. Similar to CL-WIFI, CL-ESA trained on Wikipedia achieve much higher precision than CL-NEA. However, the more coarse-grained annotation of CL-ESA yields more correct links but slightly lower precision than CL-WIFI.

In summary, our experiments show that there are significant differences regarding the performance of the investigated approaches. As reasons we indicate the different training methods (using Wikipedia data or feature sets) and linking style (fine-grained or coarse-grained). Furthermore, the gap between cross-lingual and monolingual annotation is quite high – more than one would expect.

4 Conclusion

In this paper, we study the problem of cross-lingual text annotation. In particular, we investigate different approaches and propose a novel framework for evaluating them based on annotation of documents extracted from a parallel corpus to Wikipedia. According to the evaluation framework, we perform experiments for an empirical comparison of different approaches w.r.t. the performance of the annotation and analyze the reason of the variation of each approach. We are not aware of any previous evaluation framework and comparison of the investigated approaches w. r. t. cross-lingual text annotation tasks, so that our work represents an important contribution to the field and provides a step towards clarifying the difference between these approaches and demonstrating their advantages and disadvantages. Since the results clearly show a significant gap between cross-lingual and monolingual annotation, we consider narrowing such gap as our future work.

Acknowledgements. The authors acknowledge the support of the European Community's Seventh Framework Programme FP7-ICT-2011-7 (XLike, Grant 288342) and of the German Federal Ministry of Education and Research (BMBF) under grant 02PJ1002 (SyncTech).

References

1. Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: Proceedings of the fifth conference on Applied natural language processing. ANLC '97, Stroudsburg, PA, USA, Association for Computational Linguistics (1997) 194–201

2. Sekine, S.: NYU: Description of the Japanese NE system used for MET-2. In: Proc. of the Seventh Message Understanding Conference (MUC-7). (1998)
3. Borthwick, A., Sterling, J., Agichtein, E., Grishman, R.: NYU: Description of the MENE Named Entity System as Used in MUC-7. In: Message Understanding Conference Proceedings MUC-7. (1998)
4. Asahara, M., Matsumoto, Y.: Japanese Named Entity extraction with redundant morphological analysis. 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, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 8–15
5. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4. CONLL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 188–191
6. Carreras, X., Màrquez, L., Padró, L.: A simple named entity extractor using AdaBoost. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4. CONLL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 152–155
7. Schapire, R.E., Singer, Y.: Improved Boosting Algorithms Using Confidence-rated Predictions. Mach. Learn. **37**(3) (December 1999) 297–336
8. Faruqui, M., Padó, S.: Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In: Proceedings of KONVENS 2010, Saarbrücken, Germany (2010)
9. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ACM (2007) 233–242
10. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management. CIKM '08, New York, NY, USA, ACM (2008) 509–518
11. Cilibrasi, R.L., Vitanyi, P.M.: The google similarity distance. Knowledge and Data Engineering, IEEE Transactions on **19**(3) (2007) 370–383
12. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on artificial intelligence. Volume 6. (2007) 12
13. Gabrilovich, E., Markovitch, S.: Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In: AAAI. (2006) 1301–1306
14. Sorg, P., Cimiano, P.: Cross-lingual Information Retrieval with Explicit Semantic Analysis. In: Working Notes of the Annual CLEF Meeting. (2008)
15. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-Based Multilingual Retrieval Model. In: Proceedings of ECIR. (2008) 522–530