Cross-Language Retrieval with Wikipedia ⋆

Péter Schönhofen András Benczúr István Bíró Károly Csalogány

Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute, Hungarian Academy of Sciences
{schonhofen, benczur, ibiro, cskaresz}@ilab.sztaki.hu

Abstract. We demonstrate a twofold use of Wikipedia for cross-lingual information retrieval. As our main contribution, we exploit Wikipedia hyperlinkage for query term disambiguation. We also use bilingual Wikipedia articles for dictionary extension. Our method is based on translation disambiguation; we combine the Wikipedia based technique with a method based on bigram statistics of pairs formed by translations of different source language terms.

1 Introduction

In the paper we describe our cross-lingual retrieval (CLIR) method used in the Ad Hoc track of CLEF 2007 [1]. Novel in our approach is the exploitation of Wikipedia hyperlink structure for query term translation disambiguation; we also use bigram statistics for disambiguation baseline. Experiments performed on 250 Hungarian and 550 German source language topics against the English target collections GH, LAT94 and LAT02 show 1–3% improvement in MAP by using Wikipedia. The MAP of translated queries was roughly 62% of the original ones for Hungarian and 80-88% for German source language queries when measured over the $TF \times IDF$-based Hungarian Academy of Sciences search engine [2].

Due to the morphological complexity, ad Hoc retrieval in Hungarian is a hard task with performances in general reported below those of the major European languages (French or Portuguese from CLEF 2006) [3–5]. Good quality overview resources on the Hungarian grammar can be found on the Wikipedia [6].

Our CLIR method falls in the branch of machine translation based on disambiguation between multiple possible dictionary entries [7]. In our method we first generate a raw word-by-word translation of the topic narrative by using the online dictionary of the Hungarian Academy of Sciences[1]. In the disambiguation phase we keep at least one translation for one word but discard off-topic translations. First we disambiguate translations of pairs of words by the bigram language model of Wikipedia, then we score remaining candidates by mapping them to Wikipedia articles and analyzing Wikipedia hyperlinks between them.

⋆ This work was supported by a Yahoo! Faculty Research Grant and by grants *MOLINGV* NKFP-2/0024/2005, NKFP-2004 project Language Miner http://nyelvbanyasz.sztaki.hu.

[1] http://dict.sztaki.hu/

When using Wikipedia linkage for query term disambiguation, our fundamental idea is to score candidate English terms based on the strength of their semantical relation to other candidates. Our method is a simplified version of translation disambiguation that typically also involves the grammatical role of the source phrase (e.g. [8]), an information unavailable for a typical query phrase.

Several researcher utilized ontologies to support disambiguation in machine translation [9], or as a base for internal representation bridging the source and target languages [10]; [11] provides an extensive theoretical discussion on this topic. However, due to the lack of ontologies which have a sufficiently wide coverage and at the same time are available in multiple languages, these methods typically construct their own ontologies through some machine learning technique over parallel corpora. Though the idea of taking advantage of Wikipedia has already emerged, either as an ontology [12] or as a parallel corpus [13], to our best knowledge, so far it has not been used for dictionary construction or to improve translation accuracy.

2 Our method

Our CLIR method consists of a word-by-word translation by dictionary, then a two-phase term translation disambiguation, first by bigram statistics, then, as our novel idea, by exploiting the Wikipedia hyperlink structure. We use the same stemming and stopword removal procedure for the target corpus, the dictionaries and Wikipedia articles; note that this may result in merging dictionary or Wikipedia entries. Stemming for English and German is performed by TreeTagger [14] and for Hungarian by an open source stemmer [4].

For dictionary we use an online dictionary as well as bilingual Wikipedia articles. The Hungarian Academy of Sciences online dictionary consists of an official and a community edited sub-dictionary, comprising of roughly 131,000 and 53,000 entries, respectively. We extend the dictionary by linked pairs of English and Hungarian Wikipedia article titles, a method that currently only slightly increases coverage since as of late 2007, there are only 60,000 articles in Hungarian Wikipedia mostly covering technical terms, geographical locations, historical events and people either not requiring translation or rarely occurring inside query text. We order translations by reliability and discard less reliable translations even if they would provide additional English translations for a source language term. The reliability order is official dictionary, community edited dictionary and finally translations generated from bilingual Wikipedia.

For German as source language we used the SMART English-German plugin version 1.4 [15] and the German and English Wiktionaries (which for many phrases specify the corresponding English and German terms, respectively, in their "Translations" section) as dictionaries. In addition, we collected translations from the titles of English and German Wikipedia articles written about exactly the same topic (and explicitly linked to each other through the cross-language facility of Wikipedia). We worked with the snapshots of Wiktionary and Wikipedia taken in September, 2007. Reliability ranking of the dictionaries is as follows: the SMART plugin (89,935 term pairs), followed by Wiktionary (9,610 new term pairs), and bilingual Wikipedia (126,091 new term pairs).

Table 1. Possible translations of Hungarian words from queries #251 (alternative medicine), #252 (European pension schemes) and #447 (politics of Pym Fortuin) along with their bigram and Wikipedia disambiguation scores. Sorting is by the combined score (not shown here).

Hungarian word	Bigram score	Wikipedia score	English word
természetes	0.0889	0.1667	natural
	0.3556	0.0167	grant
	0.0000	0.1833	natural ventilation
	0.0000	0.0167	naturalism
	0.0000	0.0167	genial
	0.0000	0.0167	naturalness
	0.0000	0.0167	artless
	0.0000	0.0010	naivete
	0.0000	0.0010	unaffected
kor[2]	0.2028	0.2083	age
	0.1084	0.1250	estate
	0.0385	0.0833	period
(kör)	0.0105	0.0625	cycle
	0.0350	0.0208	era
	0.0000	0.0625	epoch
	0.0000	0.0052	asl
	0.0000	0.0010	temp
ellentmondás	0.0090	0.1961	paradox
	0.0601	0.0589	conflict
	0.0060	0.0392	contradiction
	0.0030	0.0196	variance
	0.0060	0.0098	discrepancy
	0.0060	0.0049	contradict
	0.0060	0.0049	inconsistency

As we can see from the example of dictionary translations for some Hungarian words shown in Table 1, the dictionary typically gives a relatively large number of possible translations, whose majority evidently belongs to the wrong concept.

First we disambiguate by forming all possible pairs E, E' of English translations of different source language terms present in the same paragraph (query title, description or narratives). We precompute bigram statistics of the target corpus. We let $\text{rank}_B(E)$ be the maximum of all bigram counts with other terms E'. See the second column of Table 1 for typical scores.

Our main new idea is the second disambiguation step that uses Wikipedia transformed into a concept network based on the assumption that reference between articles indicates semantic relation. As opposed to proper ontologies such as WordNet or OpenCyc, here relations have no type (however, several researchers worked out techniques to rectify this omission, for instance [16]). Note however that Wikipedia itself is insufficient for CLIR itself since it deals primarily with complex concepts while basic nouns

[2] Term "kor" has different meanings (age, illness, cycle, heart suit) depending on the diacritics (see in brackets).

and verbs (e.g. “read”, “day”) are missing and hence we use it in combination with bigrams.

We preprocess Wikipedia in a way described in [17]; there the way special pages such as redirects, category pages etc. are handled is described in detail. We used the Wikipedia snapshot taken in August of 2006 with (after preprocessing) 2,184,226 different titles corresponding to 1,728,641 documents.

We label query terms by Wikipedia documents by Algorithm 1. First we find Wikipedia title words in the query; for multiword titles we require an exact matching sequence of the translated topic definition. In this way we obtain a set W_E of Wikipedia titles corresponding to translated words E. The final labels arise as the top ranked concepts after a ranking procedure that measures connectivity within the graph of the concept network as in Algorithm 1. In the algorithm first we rank Wikipedia documents D by the number of links to terms O in the source language, i.e. the number of such O with a translation E' that has a $D' \in W_{E'}$ linked to D. For each translation E we then take the maximum of the ranks within W_E. We add these ranks up in the case of multiword translations.

Algorithm 1 Outline of the labeling algorithm

for all English translation words E **do**
  for all Wikipedia documents D with title T_D containing E **do**
    if T_D appears as a sequence around E **then**
      add D to the list W_E
for all translation words E **do**
  for all Wikipedia documents $D \in W_E$ **do**
    $\text{rank}_W(D) \leftarrow |\{O : O$ is a source language word such that there is a translation E' and a $D' \in W_{E'}$ with a link between D and D' in Wikipedia$\}|$
  $\text{rank}(E) \leftarrow \max\{rank(D) : D \in W_E\}$

The third column of Table 1 shows scores of various English translation candidates computed from the degree of linkage between their corresponding Wikipedia article(s) and those of other candidates.

Finally we build the query based on the bigram and Wikipedia based ranks rank_B and rank_E of the individual translations, by also taking the quality $q(E)$ of the dictionary that contains the translation into account. We choose the translation that maximizes the expression below; in cases of ties we keep both:

$$q(E) \cdot (\log(\text{rank}_B(E)) + \alpha \log(\text{rank}_W(E))). \quad (1)$$

In Table 1, translations are ordered according to their combined scores. Note that for the first word, neither bigram statistics, nor Wikipedia would rank the correct translation to the first place, but their combined score does. For the second word, both scoring would select the same candidate as the best one, and for the third, Wikipedia scoring is right while the bigram statistics is wrong.

Table 2. Topics used in the CLIR experiments.

source language	English	German	Hungarian
GH	141–350	141–350	251–350
LAT 94	1–350	1–200, 250–350	251–350
LAT 02	401–450		401–450

3 Search engine

We use the Hungarian Academy of Sciences search engine [2] as our information retrieval system that uses a $TF \times IDF$-based ranking in a weighted combination of the following factors:

- Proximity of query terms as in [18, 19];
- Document length normalization [20];
- Different weights to different parts of the document (title, or location as in case of ImageCLEF-Photo topics);
- Total weight of query terms in the document; the original query is considered as a weighted OR query with reduced weights to words in description and narrative.
- The proportion of the document between the first and last query term, a value almost 1 if the document contains query terms at the beginning and at the end, and 1 / $size$ for a single occurrence.

We observed that we get the best result if the weight of the number of query terms is much higher than the $TF \times IDF$ score. In other words, we rank documents with respect to the number of query terms found inside their text, then use the $TF \times IDF$-based measurement to differentiate between documents carrying the same number of query terms.

We translate all of title, description and narrative that we all use for recognizing the concept of the query mapped into Wikipedia. For retrieval we then use translations with weights 1 for title, 0.33 for description and 0.25 for narrative.

4 Results

Retrieval performance for Hungarian as source language for the CLEF 2007 topics 401–450 is shown in Table 3. In addition we use a wide range of topics listed in Table 2 to show average performance for both Hungarian and German as source language in Table 4. We use the English topics as baseline; we also show performance for the bigram based and the combined bigram and Wikipedia translation disambiguation steps.

Table 5 shows (the first four words of) sample queries where the Wikipedia-enhanced translation is much less or much more effective than the official English translation. Mistakes corrected by Wikipedia-based disambiguation can be classified in five main groups:

- translated words are the same but are in a different grammatical form (see for example “human cloning” vs. “human clone” in topic 408);

Table 3. Performance over Hungarian topics of Table 2, including CLEF 2007 topics (LAT 02)

Corpus	method	P @ 5	R @ 5	P @ 10	R @ 10	MRR	MAP
GH	English topics	34.46	21.70	28.43	30.62	0.5304	0.2935
	Bigram	21.20	10.40	20.24	18.74	0.3857	0.1666
	Bigram + Wikipedia	23.86	12.94	21.45	20.98	0.4038	0.1826
LAT 94	English topics	33.90	16.81	28.74	24.00	0.5245	0.2638
	Bigram	20.21	11.55	17.26	16.74	0.3632	0.1515
	Bigram + Wikipedia	20.00	12.98	18.32	18.74	0.3830	0.1693
LAT 02	English topics	42.80	13.61	36.80	20.20	0.6167	0.2951
	Bigram	27.20	9.55	22.80	13.68	0.4518	0.1566
	Bigram + Wikipedia	31.60	10.68	28.20	15.95	0.5348	0.1956

Table 4. Performance over German topics of Table 2

Corpus	method	P @ 5	R @ 5	P @ 10	R @ 10	MRR	MAP
GH	English topics	34.12	27.78	27.00	36.66	0.5594	0.3537
	Bigram	28.00	24.21	22.71	32.88	0.4940	0.3011
	Bigram + Wikipedia	29.41	25.99	22.88	33.62	0.5078	0.3188
LAT-94	English topics	36.10	19.65	30.04	27.97	0.5746	0.2974
	Bigram	27.56	15.36	22.80	22.29	0.4667	0.2055
	Bigram + Wikipedia	29.43	18.34	24.39	24.98	0.4863	0.2327

- translated words are synonyms of the original ones (e.g. “Australian” vs. the informal “Aussie” in topic 407);
- insufficient information to properly disambiguate (topic 409);
- the Hungarian stemmer failed to determine the stem of a word (leaving the Hungarian word untranslated for “drug” in topic 415);
- the dictionary failed to provide any translation for a given word (see for instance “weapon” in topic 410, which should have been recognized indirectly from a Hungarian compound term).

Wikipedia based translations are usually more verbose than raw translations by introducing synonyms (like in topic 412) but also sometimes strange words (such as in topic 414). We often reintroduce important keywords lost in bigram based disambiguation as e.g. “cancer” in topic 438, “priest” in topic 433. As a result, though precision at 5 retrieved documents were only 27.20% for raw translations, a fraction of the 42.80% observed when using official English translations, Wikipedia post-processing managed to increase precision to 31.60%. Figure 1 shows the precision–recall curve as well as how the bigram and Wikipedia based disambiguation combination weight factor α as in (1) affects average precision.

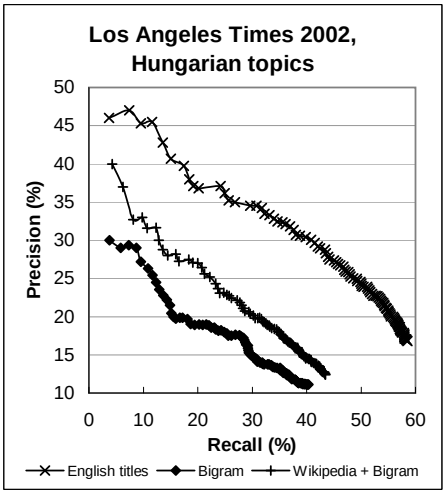
References

1. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 Ad Hoc track overview. In this volume (2007)

Table 5. Sample queries where Wikipedia based disambiguation resulted in improved (above) and deteriorated (below) average precision over bigram based disambiguation, sorted by difference.

Topic No.	Avg. prec. (En. titles)	Avg. prec. (Wikiped.)	Difference	English title	Wikipedia-enhanced translation
404	0.0667	1.0000	-0.9333	nato summit security	safety security summit nato ...
408	0.0890	0.4005	-0.3115	human cloning	human clone number statement ...
412	0.0640	0.2077	-0.1437	book politician	book politician collection anecdote ...
409	0.3654	0.4997	-0.1343	bali car bombing	car bomb bali indonesia ...
421	0.3667	0.4824	-0.1157	kostelic olympic medal	olympic kostelic pendant coin ...
438	0.0071	0.0815	-0.0744	cancer research	oncology prevention cancer treatment ...
425	0.0333	0.1006	-0.0673	endangered species	endanger species illegal slaughter ...
406	0.0720	0.1314	-0.0594	animate cartoon	cartoon award animation score ...
432	0.3130	0.3625	-0.0495	zimbabwe presidential election	presidential zimbabwe marcius victor ...
417	0.0037	0.0428	-0.0391	airplane hijacking	aircraft hijack diversion airline ...

Topic No.	Avg. prec. (En. titles)	Avg. prec. (Wikiped.)	Difference	English title	Wikipedia-enhanced translation
407	0.7271	0.0109	0.7162	australian prime minister	premier aussie prime minister ...
448	0.7929	0.1691	0.6238	nobel prize chemistry	nobel chemistry award academic ...
416	0.6038	0.0000	0.6038	moscow theatre hostage crisis	moscow hostage crisis theatre ...
410	0.6437	0.0947	0.5490	north korea nuclear weapon ...	north korean korea obligation ...
402	0.4775	0.0056	0.4719	renewable energy source	energy parent reform current ...
414	0.4527	0.0000	0.4527	beer festival	hop festival good line ...
441	0.6323	0.1974	0.4349	space tourist	tourist space russian candidate ...
443	0.5022	0.1185	0.3837	world swimming record	swim time high sport ...
427	0.6335	0.2510	0.3825	testimony milosevic	milosevic testimony versus hague ...
419	0.3979	0.0233	0.3746	nuclear waste repository	waste atom cemetery federal ...
401	0.4251	0.0899	0.3352	euro inflation	price rise euro introduction ...



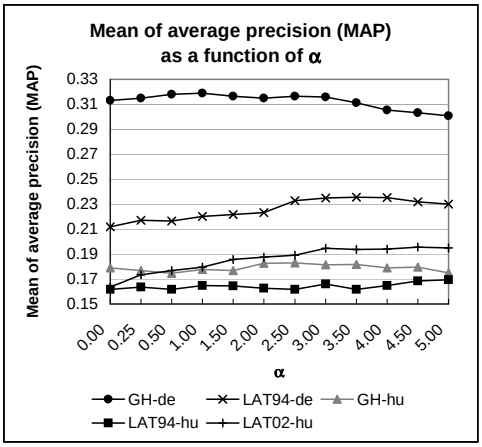


Fig. 1. Left: Precision–recall curve for the CLEF 2007 Hungarian topics when using the original English titles as well as bigram based and combined disambiguation. **Right:** Effect of α on the mean of average precision (MAP)

2. Benczúr, A.A., Csalogány, K., Fogaras, D., Friedman, E., Sarlás, T., Uher, M., Windhager, E.: Searching a small national domain – A preliminary report. In: Proceedings of the 12th International World Wide Web Conference (WWW). (2003)
3. Di Nunzio, G., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: Ad Hoc Track Overview. In CLEF 2006 proceedings, Lecture Notes in Computer Science Volume 4730, Springer (2007)
4. Halácsy, P., Trón, V.: Benefits of deep NLP-based lemmatization for information retrieval. In CLEF 2006 proceedings, Lecture Notes in Computer Science Volume 4730, Springer (2007)
5. Savoy, J., Abdou, S.: UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval. In CLEF 2006 proceedings, Lecture Notes in Computer Science Volume 4730, Springer (2007)
6. Hungarian Grammar: From Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Hungarian_grammar.
7. Hiemstra, D., de Jong, F.: Disambiguation strategies for cross-language information retrieval. In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, London, UK (1999) 274–293
8. Dorr, B.J.: The use of lexical semantics in interlingual machine translation. Machine Translation **7**(3) (1992) 135–193
9. Knight, K., Luk, S.K.: Building a large-scale knowledge base for machine translation. In: Proceedings of the twelfth National Conference on Artificial Intelligence. (1994) 773–778
10. Navigli, R., Velardi, P., Gangemi, A.: Ontology learning and its application to automated terminology translation. IEEE Intelligent Systems **18**(1) (2003) 22–31
11. Mahesh, K.: Ontology development for machine translation: Ideology and methodology. Technical Report MCCS 96-292, Computing Research Laboratory, New Mexico State University (1996)
12. Denoyer, L., Gallinari, P.: The Wikipedia XML corpus. SIGIR Forum **40**(1) (2006) 64–69
13. Adafre, S.F., de Rijke, M.: Finding similar sentences across multiple languages in Wikipedia. In: Proceedings of the New Text Workshop, 11th Conference of the European Chapter of the Association for Computational Linguistics. (2006)
14. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing. (1994)
15. Project jDictionary: SMART English-German plugin version 1.4 Available at http://jdictionary.sourceforge.net/plugins.html.
16. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipedia. In: Proceedings of the 15th international conference on World Wide Web. (2006) 585–594
17. Schönhofen, P.: Identifying document topics using the Wikipedia category network. In: Web Intelligence. (2006) 456–462
18. Rasolofo, Y., Savoy, J.: Term proximity scoring for keyword-based retrieval systems. In: ECIR. (2003) 207–218
19. Büttcher, S., Clarke, C.L.A., Lushman, B.: Term proximity scoring for Ad-Hoc retrieval on very large text collections. In: SIGIR' 06, New York, NY, ACM Press (2006) 621–622
20. Singhal, A., Buckley, C., Mitra, M., Salton, G.: Pivoted document length normalization. Technical Report TR95-1560, Cornell University, Ithaca, NY (1995)