# Performing 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 of the Hungarian Academy of Sciences {schonhofen, benczur, ibiro, cskaresz}@ilab.sztaki.hu

#### Abstract

We describe a method which is able to translate queries extended by narrative information from one language to another, with help of an appropriate machine readable dictionary and the Wikipedia on-line encyclopedia. Processing occurs in three steps: first, we look up possible translations phrase by phrase using both the dictionary and the cross-lingual links provided by Wikipedia; second, improbable translations, detected by a simple language model computed over a large corpus of documents written in the target language, are eliminated; and finally, further filtering is applied by matching Wikipedia concepts against the query narrative and removing translations not related to the overall query topic. Experiments performed on the Los Angeles Times 2002 corpus, translating from Hungarian to English showed that while queries generated at end of the second step were roughly only half as effective as original queries, primarily due to the limitations of our tools, after the third step precision improved significantly, reaching 60% of the native English level.

### **Categories and Subject Descriptors**

H.3.3 [Information Search and Retrieval]: [Query Formulation]

#### **General Terms**

Design, Experimentation, Performance

### Keywords

Cross-language information retrieval, Wikipedia

# 1 Introduction

Although the overwhelming majority of documents available on the Internet are written in English (as of 2002 roughly 72%, according to statistics presented in [23], merely a fraction of users is proficient in this language and as a consequence are most probably unable to formulate effective queries for traditional search engines. However, because they might be interested in various non-textual content stored in these documents, like diagrams, images, numeric data, hyperlinks etc., there should be a way to translate queries from any language to English. In the paper we introduce a relatively simple method which, when given a short description of the topic in Hungarian, tries

<sup>\*</sup>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.

to find the English keywords most accurately representing it. Experiments performed on the Los Angeles Times 2002 news article collection showed that precision of translated queries (computed for the top 5 retrieved documents) were roughly 60% of the original ones. This value is rather abismal, but when considering the small size of the English-Hungarian dictionary at our disposal and also the naive technique applied for raw translation, is in fact acceptible.

Our main goal was not to develop a highly accurate machine translation system, but rather to prove that employing an ontology mined from Wikipedia, quality of translations can be significantly improved. The fundamental idea is: if we score candidate English terms based on to how many other terms they are semantically related and exactly how strongly, we can reliably decide whether to retain it or not. For instance when encountering word "ég", which could be interpreted as either "sky" or "burn", and by examining the overall topic suggested by other words we found that the query is about firefighting, we can safely prefer the latter meaning. In addition, because many Wikipedia articles exist in more than one language, Hungarian and English titles of such parallel articles can yield term translations not included in a traditional dictionary.

The proposed algorithm consists of four stages. First of all, we pre-process source documents, carrying out the usual steps, such as stemming and stopword removal; in addition, we extract both phrase translations and a basic ontology from the English Wikipedia. Next, we generate a raw translation of the query narrative, utilizing a simple machine readable dictionary (SZTAKI Szótár [10]) by looking up English versions of Hungarian terms, naively word for word, then we perform disambiguation where necessary relying on a bigram language model computed over the textual contents of English Wikipedia articles – but of course any large corpus with a wide coverage would do. In the third stage, resulting terms are aligned with Wikipedia articles (in fact representing concepts), which in turn are scored primarily according to how many other recognized articles they refer to. Finally, we construct a query from titles of the few highest ranked Wikipedia articles, submitting it to a search engine, which retrieves relevant documents.

Let us examine the basic idea behind the second and third step more closely. During raw translation the dictionary often gives multiple translations for a specific Hungarian words, each valid, but only one of them is correct in the current context. Context can be represented either by simply the set of words present inside the same sentence (or paragraph), or by the set of Wikipedia concepts recognized through them. In the former case, a bigram language model might predict how "compatible" are words with each other, while in the latter case semantic connections between concepts (extracted from references between Wikipedia articles, each representing a concept described by their titles) may help us pinpoint erroneous translations.

The search engine is the Hungarian Academy of Sciences search engine [2] which uses a  $TF \times IDF$ -based ranking; it was slightly modified in order to work with the Los Angeles Times 2002 corpus and to be able to influence its ranking function in more aspects than was originally possible.

There are four research areas strongly relevant to the proposed algorithm, of course the most crucial of whose is machine translation (for an overview see for instance [16]). Machine translation is usually performed in three different ways. In the first approach, source text is converted to some kind of internal representation, typically with the help of natural language understanding, then converted to the target language (see [21], among others). In the second, a dictionary is used, and in case of ambiguous translations, the correct one is selected by determining the precise grammatical role of the source phrase (e. g. [9]). Finally, in the third, word correlation data collected from statistical analysis of parallel corpora are utilized to select the most appropriate translation (as described for example in [3]). Our algorithm adopts a simplistic solution, where neither training data acquired from parallel corpora, nor part of speech information are exploited; however, it perfectly fits our declared purpose to demonstrate how Wikipedia can improve translation quality.

Another related field is language modeling (an excellent overview is provided by [12]), and applying language modeling to machine translation [4, 6] or information retrieval [7] is not new. Though we utilize bigrams to help disambiguate English translations, we do not exploit all possibilities inherent in this technology, such as smoothing or larger *n*-grams, neither do we employ it as extensively as would be possible, for instance deciding whether a Wikipedia concept is rightly aligned to an English word sequence – based on the similarity between its immediate context in the query text and content of the corresponding article body – or not. Several researcher utilized ontologies to support disambiguation in machine translation [17], or as a base for internal representation bridging the source and target languages [20]; [19] provides an extensive theoretical discussion on this topic. However, due to a dearth 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 run over parallel corpora. Though the idea of taking advantage of Wikipedia has already emerged, either as an ontology [8] or as a parallel corpus [1], to our best knowledge, so far it has not been used for dictionary construction or to improve translation accuracy.

The last related area is cross-language information retrieval itself (see for instance [22] for an overview). There are several papers dealing with the various problematic points of the task, like acquisition of the dictionary [14], mapping terms (or concepts) from source to target languages [11, 18], or disambiguation between multiple possible translations [15]. Though our approach contains a rather simple disambiguation strategy, the main focus is on the improvement of translation accuracy via the discovery of semantic relations between recognized concepts, which is (at least presently) a quite unexplored research direction.

# 2 Preparations

Several preparational steps are necessary before we can carry out actual query translation. First of all, we have to construct a machine readable dictionary which will provide us possible English variants of individual Hungarian words or phrases. Next, a concept network should be extracted from Wikipedia, which will help us discover semantic relationships between candidate English terms, and therefore decide whether they are compatible with the query topic or not. Finally, texts of queries and documents themselves which have to be searched are pre-processed in traditional fashion, to facilitate efficient matching of its contents against dictionary entries. The next three subsection describe each above mentioned activity in detail.

### 2.1 Constructing the dictionary

It goes without saying that the most crucial component for every translation program is a machine readable dictionary, which ideally is accurate, covers a wide range of subjects, and in case of words with many possible translations also contains information about contexts characteristic to each translation. We took advantage of the collaboratively edited SZTAKI Szótár dictionary, a simple textual database of English-Hungarian word pairs behind the popular and freely available on-line service of the same name. It has two components: an official and an unofficial sub-dictionary (the main difference between them is that the former is more closely supervised), comprising of roughly 131,000 and 53,000 entries, respectively. Its structure is fairly straightforward, as can be seen from the short extracted fragment below:

```
aeronautical
repülési
```

aeronomy felsőlégkör tudománya

aeroplane repülőgép

aesthete esztéta

Obviously, the first line of each entry contains an English term, the second its Hungarian variant, while the third remains empty, serving as a very easy to detect delimiter. Unfortunately, some entries specify more than one English or Hungarian terms (typically for words having an irregular plural form, like "adman") separated by commas or semicolons; certain entries refer to idioms (for instance "blow high, blow low"), often complete with punctuations; and a small number of entries represent prefixes or postfixes, not regular phrases (e. g. "-able", "thermo-" etc). While these special dictionary items usually do not facilitate better translation accuracy, they do make our pre-processing mechanism more complex.

So when going through the original dictionary entries, we remove text inside parentheses or after a slash; discard question marks, exclamation marks and double quotes; ignore prefix or suffix references (characterized by an initial or ending hyphen); in addition, English stopwords and dictionary abbreviations (like "sb" representing a person object or subject) are deleted. When an entry contains multiple terms, it is exploded to as many entries as necessary to represent all possible English-Hungarian term pairs. For example, a compound translation from "A, B" to "C, D" will yield four entries, namely A to C, A to D, B to C and B to D.

Note that at this point we do not perform stemming neither on English, nor on Hungarian words – the former will be carried out in a subsequent processing step, and the latter is not strictly required, because terms already appear in their base forms in the dictionary. However, English words not present in any Wikipedia article bodies are deleted, partly as they are totally useless when we attach Wikipedia concepts to translated queries, partly since probably they are misspelled or extremely infrequent in common English, therefore it is very unlikely that submitting them as a query to a search engine would retrieve even a small number of documents.

Compared to professional dictionaries SZTAKI Szótár is relatively small, thus it was perfectly natural that we attempted to extend it by all available means. Luckily, articles in the English Wikipedia contain links to articles on the same subject in the Hungarian version (if present), so titles connected in this fashion can be considered valid translation pairs and added to the basic dictionary. However, there are at least two problems with the above mentioned approach: First, the number of articles in Hungarian Wikipedia is rather low, around 60,000 as of early 2007 (less than one tenth of the size of its English counterpart), a significant percentage of which is specific to the Hungarian culture, and consequently is not part of English Wikipedia. Second, Wikipedia covers mostly technical terms, geographical locations, historical events and persons, phrases either not requiring translation or rarely occurring inside common texts.

We converted cross-language references extracted from Wikipedia to dictionary entries in the following way. Similarly to the processing performed on SZTAKI Szótár, text inside parentheses or written after a comma, semicolon, slash is deleted, both in Hungarian and English titles. In addition, English titles are stemmed and from them stopwords (using a list originally developed for the Smart search engine) are removed. Note that now there is no need to break up English-Hungarian pairs, because overwhelming majority of Wikipedia articles discuss only a single concept, the rest being disambiguation pages, where the various concepts are represented by the same term, for example "Java" or "acid test", and either are translated to the same Hungarian term, or do not have a corresponding article in Hungarian Wikipedia.

After dictionary entries have been collected from both SZTAKI Szótár and Wikipedia, we merge them to produce the machine readable dictionary utilized during query translation later. As we consider entries derived from Wikipedia less reliable than those extracted from SZTAKI Szótár (which is perfectly understandable since Wikipedia originally was not intended to be employed as a dictionary), we ignore Wikipedia pairs already present among SZTAKI Szótár entries – even if they would provide additional English translations for an already known Hungarian term. It is now that Hungarian titles are stemmed and stopwords are removed (with the help of a manually constructed stopword list), making Hungarian terms of several dictionary entries the same (e. g. "futott" and "futó" are both converted to "fut"), which therefore should be merged. The resulting dictionary contains 100,510 Hungarian terms, in average composed of 1.6211 words and referring to 2.0695 English terms, each one representing a different possible sense.

#### 2.2 Converting Wikipedia to a concept network

In its original form, Wikipedia consists of articles of various types. Regular articles have a title, body, enumeration of Wikipedia categories they pertain to, and possibly links to the same article

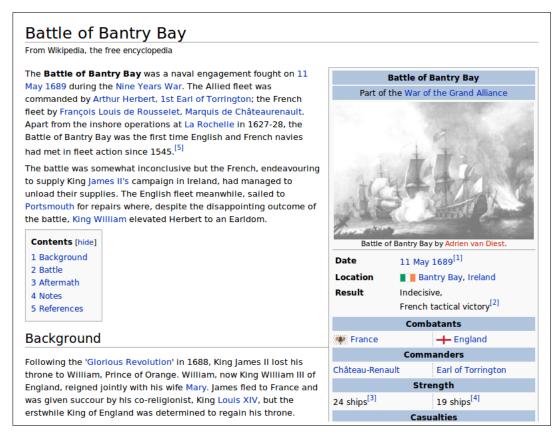


Figure 1: Fragment of a Wikipedia article.

written in other languages. Article bodies might contain supplementary tables, images, messages for contributors (such as that the article has to be reviewed or cannot be modified temporarily due to recent acts of vandalism), references to related Wikipedia articles, hyperlinks to external websites. Longer articles are usually partitioned into sections, even subsections, and consequently after a short introduction, they typically embed a table of contents. Figure 1 shows a sample.

Besides regular articles, there are two additional article types – redirects and category pages. Redirects are merely "placeholders" for alternate titles of a given article (e. g. "Eugene Paul Wigner" to "Eugene Wigner"), they have only a title without any body, category assignment list or links to versions in other languages. As their name suggests, when a user retrieves them through the official Wikipedia web interface, he or she immediately gets redirected to the appropriate regular article. Category pages describe Wikipedia categories by giving a brief overview about their exact meaning, enumerating their sub- and supercategories (if exist), finally listing all pages pertaining to them. Note that although categories are organized into a hierarchy, unfortunately it is not a tree, because categories might have multiple supercategories.

Of course, in its original form, Wikipedia cannot be used to help us recognize concepts in the English queries, then detect and discard those which are not consistent with the dominant query subject, since they have been inserted due to some erroneous judgment made by the raw translator. In fact we need only a concept network, or more precisely, an undirected graph where nodes represent concepts and edges the various relationships between them: concept names can be easily extracted from Wikipedia article titles, and if an article refers to another article, it is a fairly reliable indicator that they are semantically connected. As opposed to WordNet, OpenCyc and other proper ontologies, here we do not record the type of relations, because this information is not present in the Wikipedia corpus in explicit form (however, several researchers worked out techniques to rectify this omission, for instance [27]).

Category:Sociology		
From Wikipedia, the free encyclopedia		
Sociology is the study of social rules and proces members of voluntary associations, professional		t only as individuals, but as
A typical textbook definition of sociology calls it t Sociology is interested in our behavior as social to of short contacts between anonymous individuals	beings; thus the sociological field of in	terest ranges from the analysis
The main article for this category is <b>Sociolog</b>	<b>yy</b> .	
Contents Top • 0-9 • A B C D E F G H I J K	LMNOPQRSTUVWXYZ	Wikimedia Commons has media related to: Society
(previous 200) (next 200)		
Subcategories		
There are 24 subcategories in this category, whic	h are shown below. More may be show	vn on subsequent pages.
[+] WikiProject Sociology	cont.	C cont.
	[+] Sociological organisations	[+] Crises
[+] Branches of sociology	[+] Sociology literature	[+] Cults
[+] Branches of sociology (interdisciplinary)	Α	[+] Cultural assimilation
<ul> <li>[+] Sociological paradigms</li> <li>[+] Sociological theories</li> </ul>	[+] Appearance	D

Figure 2: Fragment of a Wikipedia category page.

To carry out the required transformations, first of all we convert bodies of regular Wikipedia articles from Wiki markup to XML, omitting embedded images, tables of contents, footnotes and other supplementary elements not strictly pertaining to the core textual content. Category pages are deleted, redirects are regarded as additional titles of articles they point to. Disambiguation pages, which list the different possible interpretations of the same term, are broken up to smaller fragments bearing the same title, otherwise our algorithm would falsely believe that articles referenced from the page all belong to a single domain, derailing the second stage of translation.

Next, titles of redirect articles are added as additional titles to the target articles, and references between articles are collected from article bodies. We apply the same pre-processing to Wikipedia article titles to which English terms inside the dictionary were subjected, in order to make future matching between query terms and Wikipedia concepts as accurate as possible. Namely we remove text between parentheses or after a comma, slash, semicolon; replace letters outside the Latin alphabet with their Latin equivalent if possible (for instance "é" with "e"); delete stopwords; and finally do stemming using TreeTagger [25].

Before we would discard Wikipedia article bodies, we split them to paragraphs and compute occurrence counts of bigrams (it goes without saying that bigrams cannot span paragraph boundaries), so that probability of word w followed by word v can be estimated with the formula:

$$P_{w,v} = P\left(v|w\right) = \frac{C_{w,v}}{f_w} \tag{1}$$

Here  $C_{w,v}$  denotes the number of occurrences of w where it was followed by v, and  $f_w$  shows the frequency of w – both calculated over the entire Wikipedia corpus. Of course, this language model could have been derived from any sufficiently large corpus; we used Wikipedia simply because it has been already converted to a format on which bigram statistics was easy to generate.

Table 1: English terms attached to query title "Kémiai Nobel-díj".

Text position	Hungarian word	Attached translations
1	kémia	chemistry
2	nobel	nobel; nobel award, nobel price, nobel prise, nobel prise, mathematics, noble prise, nobelprize, nobelprizeorg, nobelist
3	díj	award, blue ribbon, charge, due, premium, prize, remuneration, stake, trophy; nobel award, nobel price, nobel prise, nobel prise mathematics, noble prise, nobelprize, nobelprizeorg, nobelist

#### 2.3 Pre-processing queries and documents

Queries for which relevant documents had to be found in the CLEF Ad-hoc task contain three parts: title, description and narrative. While titles consist of only a few words, descriptions are whole sentences and narratives are usually even longer, sometimes being composed from two or three sentences. Though we will search in the document collection utilizing only the words captured from query titles, we translate all three query parts, otherwise there would not be a sufficient amount of context (enough recognized concept) for reliably detect query subjects. As a consequence, pre-processing of course should extend to the whole query, not just to their titles.

The pre-processing steps applied to queries do not contain any novel elements, we employ almost the same technique as for Wikipedia titles. More precisely, we perform stemming, remove stopwords (again relying on a slightly modified stopword list originally developed for the Smart search engine), if possible, convert accented characters to their equivalent in the Latin alphabet, and in addition split both query descriptions and narratives to paragraphs. Words not present in Wikipedia article bodies are ignored, because it is guaranteed that they cannot be used to recognize Wikipedia concepts and so improve translation quality.

For documents in the Los Angeles Times 2002 corpus, from which we want later to retrieve items relevant to queries, exactly the same steps are carried out, otherwise alignment between query and document content would not be as precise as would be possible.

# 3 Raw translation

As was already mentioned in the introduction of this paper, translation is carried out in two stages: first, we make a raw translation relying only on the machine readable dictionary built from SZTAKI Szótár and Wikipedia cross-language links; second, we refine the raw translation by removing terms whose meaning is not compatible with the detected subject of our query. Raw translation again consists of two stages – after we gathered possible English versions of Hungarian words, we select those seeming the most probable based on the bigram language model computed on Wikipedia article bodies, as has been described in the previous section.

So we start by examining word sequences present in the Hungarian query which we want to translate; sequences obviously have to be continuous, cannot cross paragraph or query part boundaries, and in addition their length cannot exceed 5 words. If a word sequence exactly matches some Hungarian term in our machine readable vocabulary, we attach the corresponding English term (or maybe terms) to text positions occupied by the Hungarian term. For instance, encountering "Kémiai Nobel-díj" (Nobel Prize in chemistry) as title of query No. 448 yields would yield the English terms shown in Table 1. Note that because "nobel díj" was correctly recognized as a multiword Hungarian phrase, its various English translations (the stranger ones, like "nobelprize" coming from Wikipedia) were registered for text positions 2 and 3 – as they are competing with translations assigned to individual words present at the same places.

As we can see from this example, the dictionary typically gives a relatively large number of

Term	Score	Term	Score
0.2500	$\operatorname{chemistry}$	0.0014	nobel prise mathematics
0.2500	nobel prizeorg	0.0012	premium
0.0314	award	0.0009	trophy
0.0157	nobel prise	0.0005	prize
0.0157	noble prise	0.0004	blue ribbon
0.0041	nobel	0.0002	stake
0.0026	charge	0.0000	awards
0.0020	due	0.0000	nobelprize
0.0020	nobel award	0.0000	nobelist
0.0020	nobel price	0.0000	remuneration

Table 2: Scores assigned to terms in query title "Kémiai Nobel-díj".

possible translations, whose majority is evidently wrong, and can be easily filtered out if we are able to detect that they are not present inside their typical context. Fortunately, language modeling provide a simple yet very efficient tool to help us in this respect, namely bigram statistics. The  $P_{w,v}$  values computed during pre-processing specifies the probability that word w is followed by word v, based on the analysis of a large corpus (roughly 1,600,000 Wikipedia article bodies). Let us now assume that query text contains words A and B, which can be translated to  $A'_1$  or  $A'_2$ , and  $B'_1$ or  $B'_2$ , respectively. If probability of  $A'_1$  and  $B'_2$  appearing as a bigram is significantly higher than those of the other three combinations, of course they should be chosen as the most probably valid translation. However, since terms being immediately near each other in the Hungarian version may be far apart in the corresponding English sentence, we have to tread carefully.

Our technique is the following. We pair each English terms present in the same sentence with each other in every imaginable combination (but naturally a term is never aligned with itself), and assign scores to them computed according to the formula:

$$S_t = \max_{u \in S \land u \neq t} \max\left\{ P_{t_{last}, u_{first}}, P_{t_{first}, u_{last}} \right\}$$
(2)

where  $t_{first}$  denotes the first word of term t, while  $t_{last}$  stands for its last word. To put it more intuitively, we characterize term t with the P value corresponding to the most probable bigram it participates in. This approach works well here since sentences are fairly short, obviously, if we would be dealing with longer ones, we should specify a maximal distance between t and u, otherwise the greatly increased amount of examined term pairs would surely introduce an unacceptably strong noise. Terms in the query title discussed above receive the scores shown in Table 2. As can be seen, highest ranked translations were "chemistry" and "nobelprizeorg" (derived from the web address nobelprize.org), while "awards", "nobelprize", "nobelist", "remuneration" did not form any meaningful pair with other terms, at least not according to Wikipedia articles.

After candidate English translations have been retrieved and scored, we can go through all text locations examining the terms attached to them, retaining only the highest ranked ones. It should be stressed that since several terms may get exactly the same score, ambiguous translations will be not necessarily resolved at this time (our goal now is merely to produce a raw translation).

Finally, let us answer the question what happens if a Hungarian words is not in the dictionary, since it is a proper name, technical term or a particular form of a verb or noun which was not properly reduced by the Hungarian stemmer [13]. There are two cases. If the word occurs in Wikipedia article bodies, it does not need any translation (the probability of a Hungarian word accidentally being the same to an English word, letter by letter, is extremely low). However, if it does not, we ignore it, as documents in the collection we will search using the current query will surely not contain it - remember that during pre-processing (see Section 2.3) we deleted words from documents which we did not encountered in Wikipedia.

- 1. recognize Wikipedia article titles in raw translation
- 2. retrieve Wikipedia articles (or concepts) behind the recognized titles
- 3. remove concepts not connected to any other concept
- 4. score concepts using the formula:  $L_c \times \frac{1}{1+M_c} \times F_c$
- 5. score words present in Wikipedia article titles (corresponding to concepts) using the formula:  $S_w = \sum_{l \in P_w \land l \in Q_c} S_c$
- 6. compose new query text

Figure 3: Outline of proposed algorithm.

# 4 Improving translation quality with Wikipedia

The quality of raw translation of a query title is often too low to directly submit it to the search engine for two main reasons: (1) there are still words whose English variants was not selected unambiguously, because their score was equal; and (2) it might happen that the concept most appropriate for retrieval is not present among concepts recognized in the query title, but rather it can be found among those connected to the query description or narrative. Thus we first map words and word sequences inside the English translation to Wikipedia concepts; score these concepts mainly based on how strongly they are related semantically to other concepts also recognized in the query; finally compose a new query extracting words from concepts having the highest scores. Outline of our algorithm is shown in Figure 3. It should be emphasized, however, that introducing mapping has also a few drawbacks besides its many advantages – Wikipedia deals primarily with complex concepts, basic nouns and verbs (e. g. "read", "day") are missing from it; and an additional conversion layer naturally introduces more noise into the processing.

Let us see now the various steps in more detail. In the first step, we look for Wikipedia article titles (representing concepts) exactly matching a word sequences in the English translation, again requiring that these sequences be continuous and do not span paragraph boundaries. Because CLEF-2007 queries rarely mention multiword terms – and if they do, the terms are mostly phrases from everyday English which are present in the machine readable dictionary –, Wikipedia concepts typically do not encompass more than one term. However, it may happen that due to the limitations of the dictionary a name of some person (for instance Richard Nixon) or company (e. g. Nestlé Foods) is translated in a word by word fashion, while Wikipedia contains an appropriate article. So we should allow that Wikipedia article titles cover many terms, but at the same time also demand that these terms immediately follow each other. This way "John Smith" will be correctly recognized as John Smith, but our algorithm will not retrieve it for "John met Smith".

Next, we collect actual Wikipedia articles, in other words concepts, behind the Wikipedia article titles discovered previously. Note that as a consequence of redirects, stemming, stopword removal and deletion of auxiliary descriptors (between parentheses or after a comma, slash, semicolon), a given title often points to several concepts – for example "power" to "Political Power", "Power (electricity)", "Power (physics)" etc). So besides ambiguity originating from raw translation, we are now saddled with ambiguity arising from title-concept mapping. Still, results discussed in Section 6 will prove that the scoring formula soon introduced is able to relatively effectively select the best candidate, no matter how high their number is.

In the third step, for each candidate concept c we determine the  $R_c$  set of other candidate concepts related to it, now not necessarily from the same paragraph. Remember that two concepts are considered semantically related if the body of either one contains a hyperlink reference to the

Table 3: Scores assigned to concepts mined from query No. 448.

Score	$\operatorname{Concept}$
3.3333	Nobel Prize
3.0000	Ludvig Nobel
3.0000	Alfred Nobel
3.0000	Academia
2.0000	$\operatorname{Chemistry}$
1.6667	Nobel Prize in Literature
1.0000	Fundraising
1.0000	Work (thermodynamics)
1.0000	$\operatorname{Employment}$
0.6667	Award 1002
0.4000	Miguel de la Espriella - "Noble"

other. However, if the two concepts have the same title, they are alternative (and competing) interpretations of the same word sequence, and obviously cannot support each other during later analysis, so this sort of connection is ignored. Concepts without any connections are discarded, as they probably do not pertain to the query subject. Still, this is not necessarily true in the reverse direction: just because a concept has at least one connection does not automatically mean that it is about the query topic, as it may be easily related to a similarly inappropriate concept.

In the fourth step, we assign scores to candidate concepts according to the following simple formula to estimate their relevance to the query subject:

$$S_c = L_c \times \frac{1}{1 + M_c} \times F_c \tag{3}$$

In order to simplify explanation of the above calculation, lets introduce the term "c is attached to text position p", meaning that p is occupied by a word being part of a word sequence through which c has been recognized in the first step.  $L_c$  denotes the number of text locations to which concepts related to c are attached;  $M_c$  is the number of concepts c competes with (some concept is a competitor if it is attached to at least one text location to which c is also attached); and finally  $F_c$  specifies the amount of text locations to which c is attached.

To put it more informally,  $L_c$  measures how strongly concept c correlates to the whole query description – the reason we deal with text positions and not with candidate concepts is that we would like to avoid Wikipedia titles attracting a high number of concepts to gain undeservedly large influence over the disambiguation process.  $M_c$  estimates reliability of c: if c has no competitors (the words it covers do not have alternative interpretations), it is almost surely correct, but if it conflicts with many other candidates, the probability that it will be chosen as the best translation is lower.  $F_c$  corresponds to the popular and widely used TF term frequency value, and thus emphasizes (or suppresses) the relevance of c not in relation to its competitors, like  $L_c$ , but rather to concepts attached to different word sequences. Table 3 shows scores of concepts attached to the content of query No. 448 about the Nobel Prize in Chemistry.

After we ranked candidate concepts, we are ready to map them back to English words in step five. However, instead of simply choosing the few highest ranked concepts and pasting their titles one after the other to form a new query, we follow a slightly more sophisticated approach, and introduce another layer of decision by scoring words according to the formula below:

$$S_w = \sum_{l \in P_w \land l \in Q_c} S_c \tag{4}$$

where  $P_w$  represents the set of text locations occupied by word w, and similarly,  $Q_c$  stands for the set of text locations to which concept c is attached. Thus in the above formula we add up scores of concepts recognized through titles containing w, as many times as the recognition took place,

Table 4:	Scores of	words	occurring	in	query	No.	448.
----------	-----------	-------	-----------	----	-------	-----	------

Score	Word
31.3333	nobel
13.3333	prise
7.0667	noble
4.0000	$\operatorname{chemistry}$
3.3333	$\operatorname{mathematics}$
3.0000	scholarly
3.0000	academic
2.0000	work
1.0000	contribution
0.6667	award

so more frequent words will automatically receive higher scores. Although members of multiword terms might get different scores, since they do not necessarily serve as base for exactly the same concepts throughout the query text, this is not always a bad thing, because more significant words (like "programming" in "programming language", assuming a query about computers) will be emphasized. Table 4 shows scores of words, assuming concepts scores shown in Table 3.

The query to be submitted for the search engine is formed in the final, sixth step. Remember that although we translated and analyzed the entire query (its title, description and narrative), we intend to select only a few words, since unfortunately our search engine presently do not cope well with long queries: individual query words cannot be supplemented with weights indicating their relative importance, thus insignificant words can easily distort the hit list. The applied selection heuristic is fairly simple: we retain words which occur in raw translation of the query title or have scores putting them among the top 3 words. In addition, words which appear in the original Hungarian query title but were not translated (since they were present neither in the machine readable dictionary nor in Wikipedia) are also kept.

# 5 Search engine

We use the Hungarian Academy of Sciences search engine [2] as our information retrieval system. Its ranking algorithm was developed originally for HTML pages; it uses a  $TF \times IDF$ -based ranking with increased weights for query words within URLs, anchor texts, the title and additional significant HTML elements. More precisely, the relevance score is computed from the weighted combination of the following factors:

- traditional  $TF \times IDF$ -based ranking which takes into account how close are the query terms to each other in documents [24, 5], utilizes document length normalization [26], and places stronger emphasis on query terms found in the document title than those appearing in the document body.
- Number of different query terms in the document.
- The coverage of the first and the last query term occurrence. This is the proportion of the document between the first and last query term. It is almost 1 if the document contains query terms at the beginning and at the end, and 1 / size if there is only one query term.

In contrast to the original version, here we do not require that a document contain all the query words in order to include it in the resulting hit list.

This ranking method had several parameters: weights of the  $TF \times IDF$  score and the number of query words and the query term coverage, weights of the title and body. After several runs with different settings, we have found 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.

# 6 Results

After the detailed description of our proposed algorithm let us see how well it performed in practice. For the experiments, we used the Wikipedia snapshot taken in August of 2006; as the result of various pre-processing procedures, it contained 1,728,641 concepts reachable through 2,184,226 different titles. The target corpus, the Los Angeles Times 2002 collection comprised of 86,002 news articles, in which documents relevant to 50 queries had to be retrieved. Each query was submitted to our search engine in three different versions:

- using words in the official English translation of query titles;
- using words in the raw translation of original Hungarian query titles;
- using words selected by our algorithm.

The first batch of queries served as a baseline; the second helped us estimate how well (or badly) raw translation worked; and the last demonstrated how much improvement could be achieved by the concept mapping and remapping utilizing Wikipedia. Table 5 shows exact texts of queries in the three batches – in order to keep the presentation as possible, in cases where a query contained more than four words, only the first four is shown, rest is represented by an ellipsis. As we can see, even the raw translation managed to provide a perfect translation for several queries, however, in the majority of cases it made mistakes, which 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 query No. 408);
- translated words are synonyms of the original ones (e. g. "Australian" vs. the more informal "Aussie" in the immediately preceding query, No. 407);
- bigram statistics derived from the Wikipedia corpus was not enough to properly disambiguate possible translations (see query No. 433);
- the Hungarian stemmer failed to determine the root form of a word (as happened in the case of "drogoz", "drug" in English, in query No. 415);
- the dictionary failed to provide any translation for a given word (see for instance "weapon" inside query No. 410, which should have been recognized indirectly from "atomsorompó").

Translations aligned with Wikipedia concepts are usually more verbose than raw translations, and frequently introduce synonyms (like in query No. 412) or sometimes even downright strange words (such as in query No. 414), probably because the query text does not provide a sufficiently large textual body to reliably determine the dominant topic. However, these query versions often bring back important keywords lost during raw translation – see for instance "cancer" in query No. 438, or "priest" in query No. 433). As a result, though precision at 5 retrieved documents were only 14.66% for raw translations, a fraction of the 33.31% observed when using official English translations, Wikipedia post-processing managed to increase precision to 16.45% (the values reflect the search engine setting which yielded the best accuracy). Figure 4 shows more detailed information about the performance.

Qry	Official translation	Raw translation	Improved translation
401	euro inflation	euro rise prices	euro rise continental
402	renewable energy source	power well	power energy owe
403	act cop	actor police role	actor police role sleuth actress
404	nato summit security	nato summit security	nato summit security north
405	childhood asthma	childhood asthma	childhood infancy development
406	animate cartoon	animated cartoon	animated cartoon festival
407	australian prime minister	aussie prime minister premier	aussie minister premier prime 2002
408	human cloning	human clone	clone human embryo medical
409	bali car bombing	bali car bomb	bali car bomb indonesian
410	north korea nuclear weapon	north korea dpr republic	north people korea democratic
411	best picture oscar	film oscar	film oscar movie
412	book politician	politic book	politic book political
413	reduce diabetes risk	diabetes hazard mitigation	hazard mitigation play form
414	beer festival	beer festival	beer festival john space
415	drug abuse	drogoz	drug dependence consumerism
416	moscow theatre hostage crisis	hostage crisis moscow theatre	hostage crisis moscow theatre send
417	airplane hijacking	air plane hijacking diversion	plane aeroplane hijacking diversion
418	bulent ecevit statement	bulent ecevit protestation statement	bulent ecevit revelation turkish
419	nuclear waste repository	atom feed cutting cemetery	atom cemetery waste
420	obesity ill health	prevailing hygienic problem	hygienic problem disease number
421	kostelic olympic medal	kostelic pendent pendant olympics	win event kostelic ivica
422	industrial business closure	factory business cessation	factory business mill shutdown
423	alternative flu shot	flu immunization alternative	alternative method detail give
424	internet banking increase	international network bank	network international bank information
425	endangered species	endangered species	illegal person activity
426	9/11 counterterrorism measure	11th sept 2001 aftermath	11 sept 2001 terrorism
427	testimony milosevic	milosevics testimony deposition	testimony deposition witness
428	ecological tourism	beloved blank safari park	park safari tourism
429	water health risk	polluted water hazard	polluted water sanitary
430	cosmetic procedure	cosmetic procedures	cosmetic surgical intervention
431	french presidential candidate	french president nominee	french president nominee party
432 433	zimbabwe presidential election child abuse priest	zimbabwean presidential divine sexually child	zimbabwean presidential 2002 latvian divine sexually child priest
433	political instability venezuela	politics doubt venezuelan	politics doubt venezuelan president
434 435	cause air pollution	air pollution ground	air pollution ground smog
435	vip divorce	vip breakup break divorce	vip break divorce actor
$430 \\ 437$	enron auditing irregularity	accounting irregularity	irregularity accounting enron auditor
437	cancer research	oncology	oncology cancer treatment
438	accident work	workplace accident	oncorogy cancer treatment
439	winter olympics doping scandal	dopping scandal winter olympics	scandal winter jock
440	space tourist	russian expedition candidate	russian expedition person cosmonaut
442	gueen mother funeral	mother queen funeral	mother queen funeral
443	world swimming record	float all time high	float cork sport meter
444	brazil world soccer champion	brasil footballer world champion	brasil footballer champion world
445	prince harry drug	harry prince drug	prince drug controversy
446	flood damage cultural heritage	flood damage cultural heritage	damage flood cultural heritage
447	pim fortuyn politics	archimedes constant fortuyn policy	constant archimedes fortuyn political
448	nobel prize chemistry	chemistry nobel award nobelprizeorg	chemistry nobel prise award
449	civil war africa	africa civil war	africa war civil african
450	failed assassination attempt	abortive attempt	abortive person target
	p -	r	r o

# Table 5: Queries submitted to the search engine.

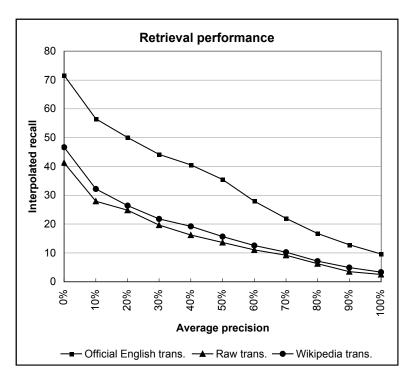


Figure 4: Interpolated recall vs. average precision when using official English translations, raw translations or Wikipedia-enhanced translations as queries.

# Conclusion and future work

In this paper we have demonstrated that a concept network extracted from the Wikipedia online encyclopedia can be used to improve the quality of machine translation of CLEF queries supplemented by narratives – thus making accurate cross-lingual information retrieval feasible. As experiments performed on the Los Angeles Times 2002 corpus have shown, our proposed algorithm works well only if it is supported by a reliable stemmer, sufficiently rich dictionary, and a relatively long text describing the purpose of the query. However, when these conditions are met, adequate precision can be achieved even with a very simple machine translation approach.

Obviously, there are several points where our method can be (and should be) improved. First of all, we plan to test it against language pairs other than Hungarian-English – German-English translation is particularly promising, as for this combination an extremely rich concept mapping is available inside Wikipedia. Second, we have to make disambiguation during raw translation more aggressive, and less dependent on the quality of stemmer which is used for the source or target language. Third, semantic relationships between Wikipedia articles could be derived from additional sources, not only from hyperlinks, such as category assignments and similarity of text of actual article bodies. Finally, we may easily re-order query results by detecting Wikipedia concepts in their text and comparing them to the concepts recognized in the query.

# References

- [1] Sisay Fissaha Adafre and Maarten de Rijke. 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.
- [2] András A. Benczúr, Károly Csalogány, Eszter Friedman, Dániel Fogaras, Tamás Sarlós, Máté Uher, and Eszter Windhager. Searching a small national domain – preliminary report. In Proceedings of the 12th World Wide Web Conference, 2003.

- [3] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.
- [4] Ralf Brown and Robert Frederking. Applying statistical english language modeling to symbolic machine translation. In Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation, pages 221–239, 1995.
- [5] Stefan Büttcher, Charles L. A. Clarke, and Brad Lushman. Term proximity scoring for ad-hoc retrieval on very large text collections. In *Proceedings of the 29th annual international ACM* SIGIR conference on Research and development in information retrieval, pages 621–622, New York, NY, 2006. ACM Press.
- [6] E. Charniak, K. Knight, and K. Yamada. Syntax-based language models for statistical machine translation. In Proceedings of the 9th Summit if the International Association for Machine Translation, 2003.
- [7] W. Bruce Croft and John Lafferty. Language Modeling for Information Retrieval. Kluwer Academic Publishers, Norwell, MA, 2003.
- [8] Ludovic Denoyer and Patrick Gallinari. The wikipedia xml corpus. SIGIR Forum, 40(1):64–69, 2006.
- Bonnie J. Dorr. The use of lexical semantics in interlingual machine translation. Machine Translation, 7(3):135-193, 1992.
- [10] Számítástechnikai és Automatizálási Kutató Intézet. Sztaki szótár. Accessible at http://szotar.sztaki.hu.
- [11] J. Gilarranz, J. Gonzalo, and F. Verdejo. An approach to conceptual text retrieval using the EuroWordNet multilingual semantic database. In Proceedings of the AAAI-96 Spring Symposium Cross-Language Text and Speech Retrieval, 1996.
- [12] J. T. Goodman. A bit of progress in language modeling. Computer Speech & Language, 15(4):403-434, 2001.
- [13] Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. A szószablya projekt. In Proceedings of the 1st Hungarian Computational Linguistics Conference, 2003.
- [14] D. Hiemstra, F. de Jong, and W. Kraaij. A domain specific lexicon acquisition tool for crosslanguage information retrieval. In Proceedings of the RIAO Conference on Computer Assisted Information Searching on Internet, pages 255–270, 1997.
- [15] Djoerd Hiemstra and Franciska de Jong. Disambiguation strategies for cross-language information retrieval. In Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, pages 274–293, London, UK, 1999.
- [16] W. John Hutchins and Harold L. Somers. An Introduction to Machine Translation. Academic Press, 1992.
- [17] Kevin Knight and Steve K. Luk. Building a large-scale knowledge base for machine translation. In Proceedings of the twelfth National Conference on Artificial Intelligence, pages 773-778, 1994.
- [18] Raija Lehtokangas, Eija Airio, and Kalervo Järvelin. Transitive dictionary translation challenges direct dictionary translation in CLIR. Information Processing and Management, 40(6):973–988, 2004.

- [19] K. Mahesh. Ontology development for machine translation: Ideology and methodology. Technical Report MCCS 96-292, Computing Research Laboratory, New Mexico State University, 1996.
- [20] Roberto Navigli, Paola Velardi, and Aldo Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1):22–31, 2003.
- [21] Sergei Nirenburg, Victor Raskin, and Allen Tucker. On knowledge-based machine translation. In Proceedings of the 11th coference on Computational linguistics, pages 627–632, Morristown, NJ, 1986. Association for Computational Linguistics.
- [22] Douglas W. Oard and Bonnie J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland at College Park, 1996.
- [23] E. T. O'Neill, B. F. Lavoie, and R. Bennett. Trends in the evolution of the public web. D-Lib Magazine, 9(4), 2003.
- [24] Yves Rasolofo and Jacques Savoy. Term proximity scoring for keyword-based retrieval systems. In Proceedings of the 25th European Conference on Information Retrieval Research, pages 207-218, 2003.
- [25] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, 1994.
- [26] Amit Singhal, Chris Buckley, Mandar Mitra, and Gerard Salton. Pivoted document length normalization. Technical Report TR95-1560, Cornell University, Ithaca, NY, 1995.
- [27] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic wikipedia. In Proceedings of the 15th international conference on World Wide Web, pages 585–594, 2006.