Translation and Extension of Concepts Across Languages

Dmitry Davidov
ICNC
The Hebrew University of Jerusalem
dmitry@alice.nc.huji.ac.il

Ari Rappoport
Institute of Computer Science
The Hebrew University of Jerusalem
arir@cs.huji.ac.il

Abstract

We present a method which, given a few words defining a concept in some language, retrieves, disambiguates and extends corresponding terms that define a similar concept in another specified language. This can be very useful for cross-lingual information retrieval and the preparation of multi-lingual lexical resources. We automatically obtain term translations from multilingual dictionaries and disambiguate them using web counts. We then retrieve web snippets with co-occurring translations, and discover additional concept terms from these snippets. Our term discovery is based on co-appearance of similar words in symmetric patterns. We evaluate our method on a set of language pairs involving 45 languages, including combinations of very dissimilar ones such as Russian, Chinese, and Hebrew for various concepts. We assess the quality of the retrieved sets using both human judgments and automatically comparing the obtained categories to corresponding English WordNet synsets.

1 Introduction

Numerous NLP tasks utilize lexical databases that incorporate concepts (or word categories): sets of terms that share a significant aspect of their meanings (e.g., terms denoting types of food, tool names, etc). These sets are useful by themselves for improvement of thesauri and dictionaries, and they are also utilized in various applications including textual entailment and question answering. Manual development of lexical databases is labor intensive, error prone, and susceptible to arbitrary human decisions. While databases like WordNet (WN) are invaluable for NLP, for some applications any offline resource would not be extensive enough. Frequently, an application requires data on some very specific topic or on very recent news-related events. In these cases even huge and ever-growing resources like Wikipedia may provide insufficient coverage. Hence applications turn to Web-based on-demand queries to obtain the desired data.

The majority of web pages are written in English and a few other salient languages, hence most of the web-based information retrieval studies are done on these languages. However, due to the substantial growth of the multilingual web[1], queries can be performed and the required information can be found in less common languages, while the query language frequently does not match the language of available information.

Thus, if we are looking for information about some lexical category where terms are given in a relatively uncommon language such as Hebrew, it is likely to find more detailed information and more category instances in a salient language such as English. To obtain such information, we need to discover a word list that represents the desired category in English. This list can be used, for instance, in subsequent focused search in order to obtain pages relevant for the given category. Thus given a few Hebrew words as a description for some category, it can be useful to obtain a similar (and probably more extended) set of English words representing the same category.

In addition, when exploring some lexical category in a common language such as English, it is

[1] http://www.internetworldstats.com/stats7.htm

frequently desired to consider available resources from different countries. Such resources are likely to be written in languages different from English. In order to obtain such resources, as before, it would be beneficial, given a concept definition in English, to obtain word lists denoting the same concept in different languages. In both cases a concept as a set of words should be translated as a whole from one language to another.

In this paper we present an algorithm that given a concept defined as a set of words in some source language discovers and extends a similar set in some specified target language. Our approach comprises three main stages. First, given a few terms, we obtain sets of their translations to the target language from multilingual dictionaries, and use web counts to select the appropriate word senses. Next, we retrieve search engine snippets with the translated terms and extract symmetric patterns that connect these terms. Finally, we use these patterns to *extend* the translated concept, by obtaining more terms from the snippets.

We performed thorough evaluation for various concepts involving 45 languages. The obtained categories were manually verified with two human judges and, when appropriate, automatically compared to corresponding English WN synsets. In all tested cases we discovered dozens of concept terms with state-of-the-art precision.

Our major contribution is a novel framework for concept translation across languages. This framework utilizes web queries together with dictionaries for translation, disambiguation and extension of given terms. While our framework relies on the existence of multilingual dictionaries, we show that even with basic 1000 word dictionaries we achieve good performance. Modest time and data requirements allow the incorporation of our method in practical applications.

In Section 2 we discuss related work, Section 3 details the algorithm, Section 4 describes the evaluation protocol and Section 5 presents our results.

2 Related work

Substantial efforts have been recently made to manually construct and interconnect WN-like databases for different languages (Pease et al., 2008; Charoenporn et al., 2007). Some studies (e.g., (Amasyali, 2005)) use semi-automated methods based on language-specific heuristics and dictionaries.

At the same time, much work has been done on automatic lexical acquisition, and in particular, on the acquisition of concepts. The two main algorithmic approaches are pattern-based discovery, and clustering of context feature vectors. The latter represents word contexts as vectors in some space and use similarity measures and automatic clustering in that space (Deerwester et al., 1990). Pereira (1993), Curran (2002) and Lin (1998) use syntactic features in the vector definition. (Pantel and Lin, 2002) improves on the latter by clustering by committee. Caraballo (1999) uses conjunction and appositive annotations in the vector representation. While a great effort has focused on improving the computational complexity of these methods (Gorman and Curran, 2006), they still remain data and computation intensive.

The current major algorithmic approach for concept acquisition is to use lexico-syntactic patterns. Patterns have been shown to produce more accurate results than feature vectors, at a lower computational cost on large corpora (Pantel et al., 2004). Since (Hearst, 1992), who used a manually prepared set of initial lexical patterns in order to acquire relationships, numerous pattern-based methods have been proposed for the discovery of concepts from seeds (Pantel et al., 2004; Davidov et al., 2007; Pasca et al., 2006). Most of these studies were done for English, while some show the applicability of their method to some other languages including Russian, Greek, Czech and French.

Many papers directly target specific applications, and build lexical resources as a side effect. Named Entity Recognition can be viewed as an instance of the concept acquisition problem where the desired categories contain words that are names of entities of a particular kind, as done in (Freitag, 2004) using co-clustering and in (Etzioni et al., 2005) using predefined pattern types. Many Information Extraction papers discover relationships between words using syntactic patterns (Riloff and Jones, 1999).

Unlike in the majority of recent studies where the acquisition framework is designed with specific languages in mind, in our task the algorithm should be able to deal well with a wide variety of target languages without any significant manual adaptations. While some of the proposed frameworks could potentially be language-independent, little research has been done to confirm it yet.

There are a few obstacles that may hinder applying common pattern-based methods to other languages. Many studies utilize parsing or POS tagging, which frequently depends on the availability and quality of language-specific tools. Most studies specify seed patterns in advance, and it is not clear whether translated patterns can work well on different languages. Also, the absence of clear word segmentation in some languages (e.g., Chinese) can make many methods inapplicable.

A few recently proposed concept acquisition methods require only a handful of seed words (Davidov et al., 2007; Pasca and Van Durme, 2008). While these studies avoid some of the obstacles above, it still remains unconfirmed whether such methods are indeed language-independent. In the concept extension part of our algorithm we adapt our concept acquisition framework (Davidov and Rappoport, 2006; Davidov et al., 2007; Davidov and Rappoport, 2008a; Davidov and Rappoport, 2008b) to suit diverse languages, including ones without explicit word segmentation. In our evaluation we confirm the applicability of the adapted methods to 45 languages.

Our study is related to cross-language information retrieval (CLIR/CLEF) frameworks. Both deal with information extracted from a set of languages. However, the majority of CLIR studies pursue different targets. One of the main CLIR goals is the retrieval of *documents* based on explicit queries, when the document language is not the query language (Volk and Buitelaar, 2002). These frameworks usually develop language-specific tools and algorithms including parsers, taggers and morphology analyzers in order to integrate multilingual *queries* and *documents* (Jagarlamudi and Kumaran, 2007). Our goal is to develop and evaluate a *language-independent* method for the translation and extension of *lexical categories*. While our goals are different from CLIR, CLIR systems can greatly benefit from our framework, since our translated categories can be directly utilized for subsequent document retrieval.

Another field indirectly related to our research is Machine Translation (MT). Many MT tasks require automated creation or improvement of dictionaries (Koehn and Knight, 2001). However, MT mainly deals with translation and disambiguation of words at the sentence or document level, while we translate whole concepts defined independently of contexts. Our primary target is not translation of given words, but the discovery and extension of a concept in a target language when the concept definition is given in some different source language.

3 Cross-lingual Concept Translation Framework

Our framework has three main stages: (1) given a set of words in a source language as definition for some concept, we automatically translate them to the target language with multilingual dictionaries, disambiguating translations using web counts; (2) we retrieve from the web snippets where these translations co-appear; (3) we apply a pattern-based concept extension algorithm for discovering additional terms from the retrieved data.

3.1 Concept words and sense selection

We start from a set of words denoting a category in a source language. Thus we may use words like *(apple, banana, ...)* as the definition of fruits or *(bear, wolf, fox, ...)* as the definition of wild animals[2]. Each of these words can be ambiguous. Multilingual dictionaries usually provide many translations, one or more for each sense. We need to select the appropriate translation for each term. In practice, some or even most of the category terms may be absent in available dictionaries. In these cases, we attempt to extract "chain" translations, i.e., if we cannot find Source→Target translation, we can still find some indirect Source→Intermediate1→Intermediate2→Target paths. Such translations are generally much more ambiguous, hence we allow up to two intermediate languages in a chain. We collect all possible translations at the chains having minimal length, and skip category terms for whom this process results in no translations.

Then we use the conjecture that terms of the same concept tend to co-appear more frequently than ones belonging to different concepts[3]. Thus,

[2] In order to reduce noise, we limit the length (in words) of multiword expressions considered as terms. To calculate this limit for a language we randomly take 100 terms from the appropriate dictionary and set a limit as $Lim_{mwe} = round(avg(length(w)))$ where $length(w)$ is the number of words in term w. For languages like Chinese without inherent word segmentation, $length(w)$ is the number of characters in w. While for many languages $Lim_{mwe} = 1$, some languages like Vietnamese usually require two words or more to express terms.

[3] Our results in this paper support this conjecture.

we select a translation of a term co-appearing most frequently with some translation of a different term of the same concept. We estimate how well translations of different terms are connected to each other. Let $C = \{C_i\}$ be the given seed words for some concept. Let $Tr(C_i, n)$ be the n-th available translation of word C_i and $Cnt(s)$ denote the web count of string s obtained by a search engine. Then we select translation $Tr(C_i)$ according to:

$$F(w_1, w_2) = \frac{Cnt(\text{“}w_1 * w_2\text{”}) \times Cnt(\text{“}w_2 * w_1\text{”})}{Cnt(w_1) \times Cnt(w_2)}$$

$$Tr(C_i) = \underset{s_i}{argmax} \left(\max_{\substack{s_j \\ j \neq i}} (F(Tr(C_i, s_i), Tr(C_j, s_j))) \right)$$

We utilize the $Yahoo$! “x * y” wildcard that allows to count only co-appearances where x and y are separated by a single word. As a result, we obtain a set of disambiguated term translations. The number of queries in this stage depends on the ambiguity of concept terms translation to the target language. Unlike many existing disambiguation methods based on statistics obtained from parallel corpora, we take a rather simplistic query-based approach. This approach is powerful (as shown in our evaluation) and only relies on a few web queries in a language independent manner.

3.2 Web mining for translation contexts

We need to restrict web mining to specific target languages. This restriction is straightforward if the alphabet or term translations are language-specific or if the search API supports restriction to this language[4]. In case where there are no such natural restrictions, we attempt to detect and add to our queries a few language-specific frequent words. Using our dictionaries, we find 1–3 of the 15 most frequent words in a desired language that are unique to that language, and we ‘and’ them with the queries to ensure selection of the proper language. While some languages as Esperanto do not satisfy any of these requirements, more than 60 languages do.

For each pair A, B of disambiguated term translations, we construct and execute the following 2 queries: {“A * B”, “B * A”}[5]. When we have 3 or more terms we also add {A B C ...}-like conjunction queries which include 3–5 terms. For languages with $Lim_{mwe} > 1$, we also construct queries with several “*” wildcards between terms. For each query we collect snippets containing text fragments of web pages. Such snippets frequently include the search terms. Since $Yahoo$! allows retrieval of up to the 1000 first results (100 in each query), we collect several thousands snippets. For most of the target languages and categories, only a few dozen queries (20 on the average) are required to obtain sufficient data. Thus the relevant data can be downloaded in seconds. This makes our approach practical for on-demand retrieval tasks.

3.3 Pattern-based extension of concept terms

First we extract from the retrieved snippets contexts where translated terms co-appear, and detect patterns where they co-appear symmetrically. Then we use the detected patterns to discover additional concept terms. In order to define word boundaries, for each target language we manually specify boundary characters such as punctuation/space symbols. This data, along with dictionaries, is the only language-specific data in our framework.

3.3.1 Meta-patterns

Following (Davidov et al., 2007) we seek symmetric patterns to retrieve concept terms. We use two meta-pattern types. First, a *Two-Slot* pattern type constructed as follows:

$$[Prefix]\ C_1\ [Infix]\ C_2\ [Postfix]$$

C_i are slots for concept terms. We allow up to Lim_{mwe} space-separated[6] words to be in a single slot. Infix may contain punctuation, spaces, and up to $Lim_{mwe} \times 4$ words. Prefix and Postfix are limited to contain punctuation characters and/or Lim_{mwe} words.

Terms of the same concept frequently co-appear in lists. To utilize this, we introduce two additional *List* pattern types[7]:

$$[Prefix]\ C_1[Infix]\ \ (C_i[Infix])+ \qquad (1)$$

$$[Infix]\ (C_i[Infix])+\ C_n\ [Postfix] \qquad (2)$$

As in (Widdows and Dorow, 2002; Davidov and Rappoport, 2006), we define a pattern graph. Nodes correspond to terms and patterns to edges. If term pair (w_1, w_2) appears in pattern P, we add nodes N_{w_1}, N_{w_2} to the graph and a directed edge $E_P(N_{w_1}, N_{w_2})$ between them.

[4] Yahoo! allows restrictions for 42 languages.

[5] These are Yahoo! queries where enclosing words in “” means searching for an exact phrase and “*” means a wildcard for exactly one arbitrary word.

[6] As before, for languages without explicit space-based word separation Lim_{mwe} limits the number of characters instead.

[7] $(X)+$ means one or more instances of X.

3.3.2 Symmetric patterns

We consider only symmetric patterns. We define a symmetric pattern as a pattern where some category terms C_i, C_j appear both in left-to-right and right-to-left order. For example, if we consider the terms $\{apple, pineapple\}$ we select a List pattern "$(one\ C_i,)+\ and\ C_n$." if we find both "one *apple*, one *pineapple*, one guava and orange." and "one watermelon, one *pineapple* and *apple*.". If no such patterns are found, we turn to a weaker definition, considering as symmetric those patterns where the same terms appear in the corpus in at least two different slots. Thus, we select a pattern "for C_1 and C_2" if we see both "for *apple* and guava," and "for orange and *apple*,".

3.3.3 Retrieving concept terms

We collect terms in two stages. First, we obtain "high-quality" core terms and then we retrieve potentially more noisy ones. In the first stage we collect all terms[8] that are bidirectionally connected to at least two different original translations, and call them *core* concept terms C_{core}. We also add the original ones as core terms. Then we detect the rest of the terms C_{rest} that appear with more different C_{core} terms than with 'out' (non-core) terms as follows:

$G_{in}(c)=\{w \in C_{core} | E(N_w, N_c) \vee E(N_c, N_w)\}$
$G_{out}(c)=\{w \notin C_{core} | E(N_w, N_c) \vee E(N_c, N_w)\}$
$C_{rest}=\{c|\ \ |G_{in}(c)|>|G_{out}(c)|\ \}$

where $E(N_a, N_b)$ correspond to existence of a graph edge denoting that translated terms a and b co-appear in a pattern in this order. Our final term set is the union of C_{core} and C_{rest}.

For the sake of simplicity, unlike in the majority of current research, we do not attempt to discover more patterns/instances iteratively by re-examining the data or re-querying the web. If we have enough data, we use windowing to improve result quality. If we obtain more than 400 snippets for some concept, we randomly divide the data into equal parts, each containing up to 400 snippets. We apply our algorithm independently to each part and select only the words that appear in more than one part.

4 Experimental Setup

We describe here the languages, concepts and dictionaries we used in our experiments.

[8]We do not consider as terms the 50 most frequent words.

4.1 Languages and categories

One of the main goals in this research is to verify that the proposed basic method can be applied to different languages unmodified. We examined a wide variety of languages and concepts. Table 3 shows a list of 45 languages used in our experiments, including west European languages, Slavic languages, Semitic languages, and diverse Asian languages.

Our concept set was based on English WN synsets, while concept definitions for evaluation were based on WN glosses. For automated evaluation we selected as categories 150 synsets/subtrees with at least 10 single-word terms in them. For manual evaluation we used a subset of 24 of these categories. In this subset we tried to select generic categories, such that no domain expert knowledge was required to check their correctness.

Ten of these categories were equal to ones used in (Widdows and Dorow, 2002; Davidov and Rappoport, 2006), which allowed us to indirectly compare to recent work. Table 1 shows these 10 concepts along with the sample terms. While the number of tested categories is still modest, it provides a good indication for the quality of our approach.

Concept	Sample terms
Musical instruments	guitar, flute, piano
Vehicles/transport	train, bus, car
Academic subjects	physics, chemistry, psychology
Body parts	hand, leg, shoulder
Food	egg, butter, bread
Clothes	pants, skirt, jacket
Tools	hammer, screwdriver, wrench
Places	park, castle, garden
Crimes	murder, theft, fraud
Diseases	rubella, measles, jaundice

Table 1: 10 of the selected categories with sample terms.

4.2 Multilingual dictionaries

We developed a set of tools for automatic access to several dictionaries. We used Wikipedia cross-language links as our main source (60%) for offline translation. These links include translation of Wikipedia terms into dozens of languages. The main advantage of using Wikipedia is its wide coverage of concepts and languages. However, one problem in using it is that it frequently encodes too specific senses and misses common ones. Thus *bear* is translated as *family Ursidae* missing its common "wild animal" sense. To overcome these

difficulties, we also used Wiktionary and complemented these offline resources with a few automated queries to several (20) online dictionaries. We start with Wikipedia definitions, then if not found, Wiktionary, and then we turn to online dictionaries.

5 Evaluation and Results

While there are numerous concept acquisition studies, no framework has been developed so far to evaluate this type of cross-lingual concept discovery, limiting our ability to perform a meaningful comparison to previous work. Fair estimation of translated concept quality is a challenging task. For most languages there are no widely accepted concept databases. Moreover, the contents of the same concept may vary across languages. Fortunately, when English is taken as a target language, the English WN allows an automated evaluation of concepts. We conducted evaluation in three different settings, mostly relying on human judges and utilizing the English WN where possible.

1. English as source language. We applied our algorithm on a subset of 24 categories using each of the 45 languages as a target language. Evaluation is done by two judges[9].

2. English as target language. All other languages served as source languages. In this case human subjects manually provided input terms for 150 concept definitions in each of the target languages using 150 selected English WN glosses. For each gloss they were requested to provide at least 2 terms. Then we ran the algorithm on these term lists. Since the obtained results were English words, we performed both manual evaluation of the 24 categories and automated comparison to the original WN data.

3. Language pairs. We created 10 different non-English language pairs for the 24 concepts. Concept definitions were the same as in (2) and manual evaluation followed the same protocol as in (1).

The absence of exhaustive term lists makes recall estimation problematic. In all cases we assess the quality of the discovered lists in terms of precision (P) and length of retrieved lists (T).

[9]For 19 of the languages, at least one judge was a native speaker. For other languages at least one of the subjects was fluent with this language.

5.1 Manual evaluation

Each discovered concept was evaluated by two judges. All judges were fluent English speakers and for each target language, at least one was a fluent speaker of this language. They were given one-line English descriptions of each category and the full lists obtained by our algorithm for each of the 24 concepts. Table 2 shows the lists obtained by our algorithm for the category described as *Relatives* (e.g., grandmother) for several language pairs including Hebrew→French and Chinese→Czech. We mixed "noise" words into each list of terms[10]. These words were automatically and randomly extracted from the same text. Subjects were required to select all words fitting the provided description. They were unaware of algorithm details and desired results. They were instructed to accept common abbreviations, alternative spellings or misspellings like yelow∈color and to accept a term as belonging to a category if at least one of its senses belongs to it, like orange∈color and orange∈fruit. They were asked to reject terms related or associated but not belonging to the target category, like tasty∉food, or that are too general, like animal∉dogs.

The first 4 columns of Table 3 show averaged results of manual evaluation for 24 categories. In the first two columns English is used as a source language and in the next pair of columns English is used as the target. In addition we display in parentheses the amount of terms added during the extension stage. We can see that for all languages, average precision (% of correct terms in concept) is above 80, and frequently above 90, and the average number of extracted terms is above 30. Internal concept quality is in line with values observed on similarly evaluated tasks for recent concept acquisition studies in English. As a baseline, only 3% of the inserted 20-40% noise words were incorrectly labeled by judges. Due to space limitation we do not show the full per-concept behavior; all medians for P and T were close to the average.

We can also observe that the majority $(> 60\%)$ of target language terms were obtained during the extension stage. Thus, even when considering translation from a rich language such as English (where given concepts frequently contain dozens of terms), most of the discovered target language terms are not discovered through translation but

[10]To reduce annotator bias, we used a different number of noise words, adding 20–40% of the original number of words.

English→Portuguese: afilhada,afilhado,amigo,avó,avô,bisavó,bisavô, bisneta,bisneto,cônjuge,cunhada,cunhado,companheiro, descendente,enteado,filha,filho,irmã,irmão,irmãos,irmãs, madrasta,madrinha,mãe,marido,mulher,namorada, namorado,neta,neto,noivo,padrasto,pai,papai,parente, prima,primo,sogra,sogro,sobrinha,sobrinho,tia,tio,vizinho
Hebrew→French: amant,ami,amie,amis,arrière-grand-mère, arrière-grand-père,beau-frère,beau-parent,beau-père,bebe, belle-fille,belle-mère,belle-soeur,bèbè,compagnon, concubin,conjoint,cousin,cousine,demi-frère,demi-soeur, épouse,époux,enfant,enfants,famille,femme,fille,fils,foyer, frère,garcon,grand-mère,grand-parent,grand-père, grands-parents,maman,mari,mère,neveu,nièce,oncle, papa,parent,père,petit-enfant,petit-fils,soeur,tante
English→Spanish: abuela,abuelo,amante,amiga,amigo,confidente,bisabuelo, cuñada,cuñado,cónyuge,esposa,esposo,espíritu,familia, familiar,hermana,hermano,hija,hijo,hijos,madre,marido, mujer,nieta,nieto,niño, novia,padre,papá,primo,sobrina, sobrino,suegra,suegro,tía,tío,tutor, viuda,viudo
Chinese→Czech: babička,bratr,brácha,chlapec,dcera,děda,dědeček,druh, kamarád,kamarádka,mama,manžel,manželka,matka, muž,otec,podnajemnik,přítelkyně, sestra,starší,strýc, strýček, syn,ségra,tchán,tchyně,teta,vnuk,vnučka,žena

Table 2: Sample of results for the Relatives concept. Note that precision is not 100% (e.g. the Portuguese set includes 'friend' and 'neighbor').

during the subsequent concept extension. In fact, brief examination shows that less than half of source language terms successfully pass translation and disambiguation stage. However, more than 80% of terms which were skipped due to lack of available translations were re-discovered in the target language during the extension stage, along with the discovery of new correct terms not existing in the given source definition.

The first two columns of Table 4 show similar results for non-English language pairs. We can see that these results are only slightly inferior to the ones involving English.

5.2 WordNet based evaluation

We applied our algorithm on 150 concepts with English used as the target language. Since we want to consider common misspellings and morphological combinations of correct terms as hits, we used a basic speller and stemmer to resolve typos and drop some English endings. The WN columns in Table 3 display P and T values for this evaluation. In most cases we obtain $> 85\%$ precision. While these results (P=87,T=17) are lower than in manual evaluation, the task is much harder due to the large number (and hence sparseness) of the utilized 150 WN categories and the incomplete nature of WN data. For the 10 categories of Table 1 used in previous work, we have obtained (P=92,T=41) which outperforms the seed-based concept acquisition of (Widdows and Dorow, 2002; Davidov and Rappoport, 2006) (P=90,T=35) on the same concepts. However, it should be noted that our task setting is substantially different since we utilize more seeds and they come from languages different from English.

5.3 Effect of dictionary size and source category size

The first stage in our framework heavily relies on the existence and quality of dictionaries, whose coverage may be insufficient. In order to check the effect of dictionary coverage on our task, we re-evaluated 10 language pairs using reduced dictionaries containing only the 1000 most frequent words. The last columns in Table 4 show evaluation results for such reduced dictionaries. Surprisingly, while we see a difference in coverage and precision, this difference is below 8%, thus even basic 1000-word dictionaries may be useful for some applications.

This may suggest that only a few correct translations are required for successful discovery of the corresponding category. Hence, even a small dictionary containing translations of the most frequent terms could be enough. In order to test this hypothesis, we re-evaluated the 10 language pairs using full dictionaries while reducing the initial concept definition to the 3 most frequent words. The results of this experiment are shown at columns 3–4 of Table 4. We can see that for most language pairs, 3 seeds were sufficient to achieve equally good results, and providing more extensive concept definitions had little effect on performance.

5.4 Variance analysis

We obtained high precision. However, we also observed high variance in the number of terms between different language pairs for the same concept. There are many possible reasons for this outcome. Below we briefly discuss some of them; detailed analysis of inter-language and inter-concept variance is a major target for future work.

Web coverage of languages is not uniform (Paolillo et al., 2005); e.g. Georgian has much less web hits than English. Indeed, we observed a correlation between reported web coverage and the number of retrieved terms. Concept coverage and

Language	English as source		English as target			
	Manual		Manual		WN	
	T[xx]	P	T[xx]	P	T	P
Arabic	29 [12]	90	41 [35]	91	17	87
Armenian	27 [21]	93	40 [32]	92	15	86
Afrikaans	40 [29]	89	51 [28]	86	19	85
Bengali	23 [18]	95	42 [34]	93	18	88
Belorussian	23 [15]	91	43 [30]	93	17	87
Bulgarian	46 [36]	85	58 [33]	87	19	83
Catalan	45 [29]	81	56 [46]	88	21	86
Chinese	47 [34]	87	56 [22]	90	22	89
Croatian	46 [26]	90	57 [35]	92	16	89
Czech	58 [40]	89	65 [39]	94	23	88
Danish	48 [35]	94	59 [38]	97	17	90
Dutch	41 [28]	92	60 [36]	94	20	88
Estonian	35 [21]	96	47 [24]	96	16	90
Finnish	34 [21]	88	47 [29]	90	19	85
French	56 [30]	89	61 [31]	93	17	87
Georgian	22 [15]	95	39 [31]	96	16	90
German	54 [32]	91	62 [34]	92	21	83
Greek	27 [16]	93	44 [30]	95	17	91
Hebrew	38 [28]	93	45 [32]	93	18	92
Hindi	30 [10]	92	46 [28]	93	16	86
Hungarian	43 [27]	90	44 [28]	93	15	87
Italian	45 [26]	89	51 [29]	88	16	81
Icelandic	27 [21]	90	39 [27]	92	15	85
Indonesian	33 [25]	96	49 [25]	95	15	90
Japanese	40 [16]	89	50 [22]	91	20	83
Kazakh	22 [14]	96	43 [36]	97	16	92
Korean	33 [15]	88	46 [29]	89	16	85
Latvian	41 [30]	92	55 [46]	90	19	83
Lithuanian	36 [26]	94	44 [35]	95	16	89
Norwegian	37 [25]	89	46 [29]	93	15	85
Persian	17 [6]	98	40 [29]	96	15	92
Polish	38 [25]	89	55 [36]	92	17	96
Portuguese	55 [34]	87	64 [33]	90	21	85
Romanian	46 [29]	93	56 [25]	96	15	91
Russian	58 [40]	91	65 [35]	92	22	84
Serbian	19 [11]	93	36 [30]	95	17	90
Slovak	32 [20]	89	56 [39]	90	15	87
Slovenian	28 [16]	94	43 [36]	95	18	89
Spanish	53 [37]	90	66 [32]	91	23	85
Swedish	52 [33]	89	62 [39]	93	16	87
Thai	26 [13]	95	41 [34]	97	16	92
Turkish	42 [33]	92	50 [25]	93	16	88
Ukrainian	47 [33]	88	54 [28]	88	16	83
Vietnamese	26 [8]	84	48 [25]	89	15	82
Urdu	27 [14]	84	42 [36]	88	14	82
Average	38 [24]	91	50 [32]	92	17	87

Table 3: Concept translation and extension results. The first column shows the 45 tested languages. **Bold** are languages evaluated with at least one native speaker. P: precision, T: number of retrieved terms. "[xx]": number of terms added during the concept extension stage. Columns 1-4 show results for manual evaluation on 24 concepts. Columns 5-6 show automated WN-based evaluation on 150 concepts. For columns 1-2 the input category is given in English, in other columns English served as the target language.

content is also different for each language. Thus, concepts involving fantasy creatures were found to have little coverage in Arabic and Hindi, and wide coverage in European languages. For vehicles, Snowmobile was detected in Finnish and Swedish while Rickshaw appears in Hindi.

Language pair Source-Target	Regular data		Reduced seed		Reduced dict.	
	T[xx]	P	T	P	T	P
Hebrew-French	43[28]	89	39	90	35	87
Arabic-Hebrew	31[24]	90	25	94	29	82
Chinese-Czech	35[29]	85	33	84	25	75
Hindi-Russian	45[33]	89	45	87	38	84
Danish-Turkish	28[20]	88	24	88	24	80
Russian-Arabic	28[18]	87	19	91	22	86
Hebrew-Russian	45[31]	92	44	89	35	84
Thai-Hebrew	28[25]	90	26	92	23	78
Finnish-Arabic	21[11]	90	14	92	16	84
Greek-Russian	48[36]	89	47	87	35	81
Average	35[26]	89	32	89	28	82

Table 4: Results for non-English pairs. P: precision, T: number of terms. "[xx]": number of terms added in the extension stage. Columns 1-2 show results for normal experiment settings, 3-4 show data for experiments where the 3 most frequent terms were used as concept definitions, 5-6 describe results for experiment with 1000-word dictionaries.

Morphology was completely neglected in this research. To co-appear in a text, terms frequently have to be in a certain form different from that shown in dictionaries. Even in English, plurals like *spoons, forks* co-appear more than *spoon, fork*. Hence dictionaries that include morphology may greatly improve the quality of our framework. We have conducted initial experiments with promising results in this direction, but we do not report them here due to space limitations.

6 Conclusions

We proposed a framework that when given a set of terms for a category in some source language uses dictionaries and the web to retrieve a similar category in a desired target language. We showed that the same pattern-based method can successfully extend dozens of different concepts for many languages with high precision. We observed that even when we have very few ambiguous translations available, the target language concept can be discovered in a fast and precise manner without relying on any language-specific preprocessing, databases or parallel corpora. The average concept total processing time, including all web requests, was below 2 minutes[11]. The short running time and the absence of language-specific requirements allow processing queries within minutes and makes it possible to apply our method to on-demand cross-language concept mining.

[11] We used a single PC with ADSL internet connection.

References

M. Fatih Amasyali, 2005. Automatic Construction of Turkish WordNet. *Signal Processing and Communications Applications Conference*.

Sharon Caraballo, 1999. Automatic Construction of a Hypernym-Labeled Noun Hierarchy from Text. *ACL '99*.

Thatsanee Charoenporn, Virach Sornlertlamvanich, Chumpol Mokarat, Hitoshi Isahara, 2008. Semi-Automatic Compilation of Asian WordNet. *Proceedings of the 14th NLP-2008, University of Tokyo, Komaba Campus, Japan.*

James R. Curran, Marc Moens, 2002. Improvements in Automatic Thesaurus Extraction. *SIGLEX '02*, 59–66.

Dmitry Davidov, Ari Rappoport, 2006. Efficient Unsupervised Discovery of Word Categories Using Symmetric Patterns and High Frequency Words. *COLING-ACL '06.*

Dmitry Davidov, Ari Rappoport, Moshe Koppel, 2007. Fully Unsupervised Discovery of Concept-Specific Relationships by Web Mining. *ACL '07*.

Dmitry Davidov, Ari Rappoport, 2008a. Unsupervised Discovery of Generic Relationships Using Pattern Clusters and its Evaluation by Automatically Generated SAT Analogy Questions. *ACL '08.*

Dmitry Davidov, Ari Rappoport, 2008b. Classification of Semantic Relationships between Nominals Using Pattern Clusters. *ACL '08.*

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, Richard Harshman, 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Info. Science*, 41(6):391–407.

Beate Dorow, Dominic Widdows, Katarina Ling, Jean-Pierre Eckmann, Danilo Sergi, Elisha Moses, 2005. Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. *MEANING '05*.

Oren Etzioni, Michael Cafarella, Doug Downey, S. Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, Alexander Yates, 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence, 165(1):91134*.

Dayne Freitag, 2004. Trained Named Entity Recognition Using Distributional lusters. *EMNLP '04*.

James Gorman , James R. Curran, 2006. Scaling Distributional Similarity to Large Corpora *COLING-ACL '06*.

Marti Hearst, 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. *COLING '92.*

Jagadeesh Jagarlamudi, A Kumaran, 2007. Cross-Lingual Information Retrieval System for Indian Languages *Working Notes for the CLEF 2007 Workshop.*

Philipp Koehn, Kevin Knight, 2001. Knowledge Sources for Word-Level Translation Models. *EMNLP '01*.

Dekang Lin, 1998. Automatic Retrieval and Clustering of Similar Words. *COLING '98*.

Margaret Matlin, 2005. *Cognition, 6th edition.* John Wiley & Sons.

Patrick Pantel, Dekang Lin, 2002. Discovering Word Senses from Text. *SIGKDD '02*.

Patrick Pantel, Deepak Ravichandran, Eduard Hovy, 2004. Towards Terascale Knowledge Acquisition. *COLING '04*.

John Paolillo, Daniel Pimienta, Daniel Prado, et al., 2005. Measuring Linguistic Diversity on the Internet. *UNESCO Institute for Statistics Montreal, Canada*.

Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, Alpa Jain, 2006. Names and Similarities on the Web: Fact Extraction in the Fast Lane. *COLING-ACL '06*.

Marius Pasca, Benjamin Van Durme, 2008. Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs. *ACL '08*.

Adam Pease, Christiane Fellbaum, Piek Vossen, 2008. Building the Global WordNet Grid. *CIL18*.

Fernando Pereira, Naftali Tishby, Lillian Lee, 1993. Distributional Clustering of English Words. *ACL '93*.

Ellen Riloff, Rosie Jones, 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *AAAI '99*.

Martin Volk, Paul Buitelaar, 2002. A Systematic Evaluation of Concept-Based Cross-Language Information Retrieval in the Medical Domain. *In: Proc. of 3rd Dutch-Belgian Information Retrieval Workshop.* Leuven.

Dominic Widdows, Beate Dorow, 2002. A Graph Model for Unsupervised Lexical Acquisition. *COLING '02*.