Enhancement of Lexical Concepts Using Cross-lingual Web Mining

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

Sets of lexical items sharing a significant aspect of their meaning (*concepts*) are fundamental in linguistics and NLP. Manual concept compilation is labor intensive, error prone and subjective. We present a web-based concept extension algorithm. Given a set of terms specifying a concept in some language, we translate them to a wide range of intermediate languages, disambiguate the translations using web counts, and discover additional concept terms using symmetric patterns. We then translate the discovered terms back into the original language, score them, and extend the original concept by adding back-translations having high scores. We evaluate our method in 3 source languages and 45 intermediate languages, using both human judgments and WordNet. In all cases, our cross-lingual algorithm significantly improves high quality concept extension.

1 Introduction

A *concept* (or lexical category) is a set of lexical items sharing a significant aspect of their meanings (e.g., types of food, tool names, etc). Concepts are fundamental in linguistics and NLP, in thesauri, dictionaries, and various applications such as textual entailment and question answering.

Great efforts have been invested in manual preparation of concept resources such as WordNet (WN). However, manual preparation is labor intensive, which means it is both costly and slow to update. Applications needing data on some very specific domain or on a recent news-related event may find such resources lacking. In addition, manual preparation is error-prone and susceptible to subjective concept membership decisions, frequently resulting in concepts whose terms do not belong to the same level of granularity[1]. As a result, there is a need to find methods for automatic improvement of concept coverage and quality.

The web is a huge up-to-date corpus covering many domains, so using it for concept extension has the potential to address the above problems. The majority of web pages are written in a few 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[2], languages in which concept terms are expressed in the most precise manner frequently do not match the language where information is needed. Moreover, representations of the same concept in different languages may complement each other.

In order to benefit from such cross-lingual information, concept acquisition systems should be able to gather concept terms from many available languages and convert them to the desired language. In this paper we present such an algorithm. Given a set of words specifying a concept in some source language, we translate them to a range of intermediate languages and disambiguate the translations using web counts. Then we discover additional concept terms using symmetric patterns and translate the discovered terms back into the original language. Finally we score the back-translations using their intermediate languages' properties, and extend the original concept by adding back-translations having high scores. The only language-specific resource required by the algorithm are multilingual dictionaries, and its processing times are very modest.

We performed thorough evaluation for 24 concepts in 3 source languages (Hebrew, English and Russian) and 45 intermediate languages. Concept definitions were taken from existing WordNet subtrees, and the obtained new terms were manually

[1] See Section 5.1.1.

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

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 852–861,
Singapore, 6-7 August 2009.

scored by human judges. In all cases we have significantly extended the original concept set with high precision. We have also performed a fully automatic evaluation with 150 concepts, showing that the algorithm can re-discover WN concepts with high precision and recall when given only partial lists as input.

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

2 Related work

One of the main goals of this paper is the extension or automated creation of lexical databases such as WN. Due to the importance of WN for NLP tasks, substantial research was done on direct or indirect automated extension of the English WN (e.g., (Snow et al., 2006)) or WN in other languages (e.g., (Vintar and Fišer, 2008)). The majority of this research was done on extending the tree structure (finding new synsets (Snow et al., 2006) or enriching WN with new relationships (Cuadros and Rigau, 2008)) rather than improving the quality of existing concept/synset nodes. Other related studies develop concept acquisition frameworks for on-demand tasks where concepts are defined by user-provided seeds or patterns (Etzioni et al., 2005; Davidov et al., 2007), or for fully unsupervised database creation where concepts are discovered from scratch (Banko et al., 2007; Davidov and Rappoport, 2006).

Some 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 concepts 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.

The two main algorithmic approaches to the problem are pattern-based concept discovery and clustering of context feature vectors. The latter approach represents word contexts as vectors in some space and uses similarity measures and automatic clustering in that space (Deerwester et al., 1990). Pereira et al.(1993), Curran and Moens (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 great effort has been made for improving the computational complexity of these methods (Gorman and Curran, 2006), they still remain data and computation intensive.

The second major algorithmic approach 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). In concept acquisition, pattern-based methods were shown to outperform LSA by a large margin (Widdows and Dorow, 2002). 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 methods to other languages, including Greek, Czech, Slovene and French.

Most of these papers attempt to discover concepts from data available in some specific language. Recently several studies have proposed to utilize a second language or several specified languages in order to extract or extend concepts (Vintar and Fišer, 2008; van der Plas and Tiedemann, 2006) or paraphrases (Bosma and Callison-Burch, 2007). However, these methods usually require the availability of parallel corpora, which limits their usefulness. Most of these methods utilize distributional measures, hence they do not possess the advantages of the pattern-based framework.

Unlike in the majority of recent studies, where the framework is designed with specific languages in mind, in our task, in order to take advantage of information from diverse languages, the algorithm should be able to deal well with a wide variety of possible intermediate languages without any manual adaptations. Relying solely on multilingual dictionaries and the web, our algorithm should be able to discover language-specific patterns and concept terms. While some of the proposed frameworks could potentially be language-independent, little research has been done to confirm this. 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 depend on the availability and quality of language-specific tools. Some 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 and no pattern pre-specification (Davidov et al., 2007; Pasca and Van Durme, 2008). While these studies avoid some of the obstacles above, it still remains open whether such methods are indeed language-independent. In the translation to intermediate languages part of our framework, we adapt the algorithms in (Davidov and Rappoport, 2006; Davidov et al., 2007) to suit diverse languages (including ones without explicit word segmentation). We also develop a method for efficient automated disambiguation and translation of terms to and from any available intermediate language.

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 and taggers in order to integrate multilingual *queries* and *documents* (Jagarlamudi and Kumaran, 2007). Our goal is to develop a *language-independent* method using cross-lingual information, for the extension and improvement of *concepts* rather than the retrieval of documents. Besides, unlike in many CLIR frameworks, intermediate languages are not specified in advance and the language of requested data is the same as the language of request, while available information may be found in many different intermediate languages.

3 The Algorithm

Our algorithm is comprised of the following stages: (1) given a set of words in a *source* language as a specification for some concept, we automatically translate them to a diverse set of *intermediate* languages, using multilingual dictionaries; (2) the translations are disambiguated using web counts; (3) for each language, we retrieve a set of web snippets where these translations co-appear and apply a pattern-based concept extension algorithm for discovering additional terms; (4) we translate the discovered terms back to the source language, and disambiguate them; (5) we score the back-translated terms using data on their behavior in the intermediate languages, and merge the sets obtained from different languages into a single one, retaining terms whose score passes a certain threshold. Stages 1-3 of the algorithm have been described in (Davidov and Rappoport, 2009), where the goal was to translate a concept given in one language to other languages. The framework presented here includes the new stages 4-5, and its goal and evaluation methods are completely different.

3.1 Concept specification and translation

We start from a set of words denoting a concept in a given source language. Thus we may use words like *(apple, banana, ...)* as the definition of the concept of fruit or *(bear, wolf, fox, ...)* as the definition of wild animals. 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 or more words to express terms.

3.2 Disambiguation of translated terms

One of the problems in utilization of multilingual information is ambiguity of translation. First, in order to apply the concept acquisition algorithm, at least some of the given concept terms must be automatically translated to each intermediate language. In order to avoid reliance on parallel corpora, which do not exist or are extremely small for most of our language pairs, we use bilingual dictionaries. Such dictionaries usually provide many translations, one or more for each sense, so this translation is inherently fuzzy. Second, once we acquire translated term lists for each intermediate language, we need to translate them back to the source language and such back-translations are also fuzzy. In both cases, we need to select the appropriate translation for each term.

While our desire would be to work with as many languages as possible, in practice, some or even

most of the concept terms may be absent from the appropriate dictionary. Such concept terms are ignored.

One way to deal with ambiguity is by applying distributional methods, usually requiring a large single-language corpus or, more frequently, parallel corpora. However, such corpora are not readily available for many languages and domains. Extracting such statistical information on-demand is also computationally demanding, limiting its usability. Hence, we take a simple but effective query-based approach. This approach, while being powerful as we show in the evaluation, only relies on a few web queries and does not rely on any language-specific resources or data.

We use the conjecture that terms of the same concept tend to co-appear more frequently than ones belonging to different concepts[3]. Thus, 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. We select a 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”,“x * * y” wildcards that allow to count only co-appearances where x and y are separated by a single word or word pair. As a result, we obtain a set of disambiguated term translations. This method is used both in order to translate from the source language to each intermediate language and to back-translate the newly discovered concept terms from the intermediate to the source language.

The number of queries in this stage depends on the ambiguity of the concept terms' translations. In order to decrease the amount of queries, if there are more than three possible senses we sort them by frequency[4] and take three senses with medium frequency. This allows us to skip the most ambiguous and rare senses without any significant effect on performance. Also, if the number of combinations is still too high (>30), we randomly sample at most 30 of the possible combinations.

3.3 Pattern-based extension of concept terms in intermediate languages

We first mine the web for contexts containing the translations. Then 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 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.

Web mining for translation contexts. In order to get language-specific data, we need to restrict web mining each time to the processed intermediate language. This restriction is straightforward if the alphabet or term translations are language-specific or if the search API supports restriction to this language[5]. 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 proper language selection. This works well for almost all languages (Esperanto being a notable exception).

For each pair A, B of disambiguated term translations, we construct and execute the following two queries: {“A * B”, “B * A”}[6]. When we have 3 or more terms we also add {A B C D}-like conjunction queries which include 3-5 words. 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 $Yahoo!$ $Boss$ allows retrieval of up to the 1000 first results (50 in each query), we collect several thousands snippets. For most of the intermediate languages, only a few dozen queries (40 on the average) are required to obtain sufficient data, and queries can be parallelized. Thus the relevant data can be downloaded

[3] Our results here support this conjecture.

[4] Frequency is estimated by web count for a given word.

[5] Yahoo! allows restriction for 42 languages.

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

in seconds. This makes our approach practical for on-demand retrieval or concept verification tasks.

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[7] 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[8]:

$$[Prefix]\ C_1\ [Infix]\ (C_i\ [Infix])+ \quad (1)$$
$$[Infix]\ (C_i\ [Infix])+\ C_n\ [Postfix] \quad (2)$$

Following (Widdows and Dorow, 2002), 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.

Symmetric patterns. We consider only symmetric patterns. We define a symmetric pattern as a pattern where some concept 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*,".

Retrieving concept terms. We collect terms in two stages. First, we obtain "high-quality" core terms and then we retrieve potentially more noisy ones. At the first stage we collect all terms[9] 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 are connected to the core stronger than to the remaining words, as follows:
$G_{in}(c)=\{w \in C_{core} | E(N_w, N_c) \vee E(N_c, N_w)\}$
$G_{out}(c)=\{w \notin C_{core} | E(N_w, N_c) \vee E(N_c, N_w)\}$
$C_{rest}=\{c | |G_{in}(c)| > |G_{out}(c)|\}$

For the sake of simplicity, we do not attempt to discover more patterns/instances iteratively by requerying 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 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.

3.4 Back-translation and disambiguation

At the concept acquisition phase of our framework we obtained sets of terms for each intermediate language, each set representing a concept. In order to be useful for the enhancement of the original concept, these terms are now back-translated to the source language. We disambiguate each back-translated term using the process described in Section 3.2. Having sets of back-translated terms for each intermediate language, our goal is to combine these into a single set.

3.5 Scoring and merging the back translations

We do this merging using the following scoring strategy, assigning for each proposed term t' in concept C the score $S(t', C)$, and selecting terms with $S(t', C) > H$ where H is a predefined threshold.

Our scoring is based on the two following considerations. First, we assume that terms extracted from more languages tend to be less noisy and language-dependent. Second, we would like to favor languages with less resources for a given concept, since noise empirically appears to be less prominent in such languages[10].

For language L and concept $C = \{t_1 \dots t_k\}$ we get a disambiguated set of translations $\{Tr(t_1, L) \dots Tr(t_k, L)\}$. We define relative lan-

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

[8] $(E)+$ means one or more instances of E.

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

[10] Preliminary experimentation, as well as the evaluation results presented in this paper, support both of these considerations.

guage frequency by

$$LFreq(L,C) = \frac{\sum_{t_i \in C}(Freq(Tr(t_i,L)))}{\sum_{L',t_i \in C}(Freq(Tr(t_i,L'))}$$

where $Freq(Tr(t_i,L))$ is a frequency of term's t_i translation to language L estimated by the number of web hits. Thus languages in which translated concept terms appear more times will get higher relative frequency, potentially indicating a greater concept translation ambiguity. Now, for each new term t' discovered through $LNum(t')$ different languages $L_1 \ldots L_{LNum(t')}$ we calculate a term score [11] $S(t',C)$:

$$S(t',C) = LNum(t') \cdot \left(1 - \sum_i LFreq(L_i,C)\right)$$

For each discovered term t', $S(t',C) \in [0, LNum(t')]$, while discovery of t' in less frequent languages will cause the score to be closer to $LNum(t')$. So terms appearing in a greater number of infrequent languages will get higher scores.

After the calculation of score for each proposed term, we retain terms whose scores are above the predefined threshold H. In our experiments we have used $H = 3$, usually meaning that acquisition of a term through 3-4 uncommon intermediate languages should be enough to accept it. The same score measure can also be used to filter out "bad" terms in an already existing concept.

4 Experimental Setup

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

4.1 Languages and concepts

One of the main goals in this research is to take advantage of concept data in every possible language. As intermediate languages, we used 45 languages including major west European languages like French or German, Slavic languages like Russian, Semitic languages as Hebrew and Arabic, and diverse Asian languages such as Chinese and Persian. To configure parameters we have used a set of 10 concepts in Russian as a development set. These concepts were not used in evaluation.

We examined a wide variety of concepts and for each of them we used all languages with available translations. Table 1 shows the resulting top 10 most utilized languages in our experiments.

[11] In this expression i runs only on languages with term t' hence the summation is not 1.

English	**Russian**	**Hebrew**
German(68%)	English(70%)	English(66%)
French(60%)	German(62%)	German(65%)
Italian(60%)	French(62%)	Italian(61%)
Portuguese(57%)	Spanish(58%)	French(59%)
Spanish(55%)	Italian(56%)	Spanish(57%)
Turkish(51%)	Portuguese(54%)	Portuguese(57%)
Russian(50%)	Korean(50%)	Korean(48%)
Korean(46%)	Turkish(49%)	Russian(43%)
Chinese(45%)	Chinese(47%)	Turkish(43%)
Czech(42%)	Polish (44%)	Czech(40%)

Table 1: The ten most utilized intermediate languages in our experiments. In parentheses we show the percentage of new terms that these languages helped discover.

We have used the English, Hebrew (Ordan and Winter, 2008) and Russian (Gelfenbeynand et al., 2003) WordNets as sources for concepts and for the automatic evaluation. Our concept set selection was based on English WN subtrees. To perform comparable experiments with Russian and Hebrew, we have selected the same subtrees in the Hebrew and Russian WN. Concept definitions given to human judges for evaluation were based on the corresponding WN glosses. For automated evaluation we selected 150 synsets/subtrees containing at least 10 single word terms (existing in all three tested languages).

For manual evaluation we used a subset of 24 of these concepts. In this subset we tried to select generic concepts manually, such that no domain expert knowledge was required to check their correctness. Ten of these concepts were identical to ones used in (Widdows and Dorow, 2002; Davidov and Rappoport, 2006), which allowed us to compare our results to recent work in case of English. Table 2 shows these 10 concepts along with the sample terms. While the number of tested concepts is not very large, 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 2: Ten of the selected concepts with sample terms.

4.2 Multilingual dictionaries

We developed tools for automatic access to a number of 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 it has is that it frequently encodes too specific senses and misses common ones (*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 automated queries to several (25) online dictionaries. We start with Wikipedia definitions, then Wiktionary, and then, if not found, we turn to online dictionaries.

5 Evaluation and Results

Potential applications of our framework include both the extension of existing lexical databases and the construction of new databases from a small set of seeds for each concept. Consequently, in our evaluation we aim to check both the ability to extend nearly complete concepts and the ability to discover most of the concept given a few seeds. Since in our current framework we extend a small subset of concepts rather than the whole database, we could not utilize application-based evaluation strategies such as performance in WSD tasks (Cuadros and Rigau, 2008).

5.1 Human judgment evaluation

In order to check how well we can extend existing concepts, we count and verify the quality of new concept terms discovered by the algorithm given complete concepts from WN. Performing an automatic evaluation of such new terms is a challenging task, since there are no exhaustive term lists available. Thus, in order to check how well newly added terms fit the concept definition, we have to use human judges.

We provided four human subjects with 24 lists of newly discovered terms, together with original concept definitions (written as descriptive natural language sentences) and asked them to rank (1-10, 10 being best) how well each of these terms fits the given definition. We have instructed judges to accept common misspellings and reject words that are too general/narrow for the provided definition.

We mixed the discovered terms with equal amounts of terms from three control sets: (1) terms from the original WN concept; (2) randomly selected WN terms; (3) terms obtained by applying the single-language concept acquisition algorithm described in Section 3.3 in the source language. Kappa inter-annotator agreement scores were above 0.6 for all tests below.

5.1.1 WordNet concept extension

The middle column of Table 3 shows the judge scores and average amount of added terms for each source language. In this case the algorithm was provided with complete term lists as concept definitions, and was requested to extend these lists. We can see that while the scores for original WN terms are not perfect (7/10), single-language and cross-lingual concept extension achieve nearly the same scores. However, the latter discovers many more new concept terms without reducing quality. The difference becomes more substantial for Hebrew, which is a resource-poor source language, heavily affecting the performance of single-language concept extension methods.

The low ranks for WN reflect the ambiguity of definition of some of its classification subtrees. Thus, for the 'body part' concept defined in WordNet as "any part of an organism such as an organ or extremity" (which is not supposed to require domain-specific knowledge to identify) low scores were given (correctly) by judges to generic terms such as tissue, system, apparatus and process (process defined in WN as "a natural prolongation or projection from a part of an organism"), positioned in WN as direct hyponyms of body parts. Low scores were also given to very specific terms like "saddle" (posterior part of the back of a domestic fowl) or very ambiguous terms like "small" (the slender part of the back).

5.1.2 Seed-based concept extension

The rightmost column of Table 3 shows similar information to the middle column, but when only the three most frequent terms from the original WN concept were given as concept definitions. We can see that even given three words as seeds, the cross-lingual framework allows to discover many new terms. Surprisingly, terms extracted by the cross-lingual framework achieve significantly higher scores not only in comparison to the single-language algorithm but also in comparison to existing WN terms. Thus while the "native" WN concept and single-language concept extension re-

sults get a score of 7/10, terms obtained by the cross-lingual framework obtain an average score of nearly 9/10.

This suggests that our cross-lingual framework can lead to better (from a human judgment point of view) assignment of terms to concepts, even in comparison to manual annotation.

	Input	
	all terms	3 terms
English		
WordNet	7.2	7.2
Random	1.8	1.8
SingleLanguage	7.0(10)	7.8(18)
Crosslingual	6.9(19)	8.8(26)
Russian		
WordNet	7.8	7.8
Random	1.9	1.9
SingleLanguage	7.4(10)	8.1(16)
Crosslingual	7.6(21)	9.0(29)
Hebrew		
WordNet	7.0	7.0
Random	1.3	1.3
SingleLanguage	6.5(4)	7.5(6)
Crosslingual	6.8(18)	8.9(24)

Table 3: Human judgment scores for concept extension in three languages (1 . . . 10, 10 is best). The WordNet, Random and SingleLanguage rows provide corresponding baselines. Average count of newly added terms are shown in parentheses. Average original WN concept size in this set was 36 for English, 32 for Russian and 27 for Hebrew.

5.2 WordNet-based evaluation

While human judgment evaluation provides a good indication for the quality of our framework, it has severe limitations. Thus terms in many concepts require domain expertise to be properly labeled. We have complemented human judgment evaluation with automated WN-based evaluation with a greater (150) number of concepts. For each of the 150 concepts, we have applied our framework on a subset of the available terms, and estimated precision and recall of the resulting term list in comparison to the original WN term list. The evaluation protocol and metrics were very similar to (Davidov and Rappoport, 2006; Widdows and Dorow, 2002) which allowed us to do indirect comparison to previous work.

Table 4 shows precision and recall for this task comparing single-language concept extension and the cross-lingual framework. We can see that in all cases, utilization of the latter greatly improves recall. It also significantly outperforms the single-language pattern-based method introduced by (Davidov and Rappoport, 2006), which achieves average precision of 79.3 on a similar set in English (in comparison to 86.7 in this study). We can also see a decrease in precision when the algorithm is provided with 50% of the concept terms as input and had to discover the remaining 50%. However, careful examination of the results shows that this decrease is due to discovery of additional correct terms not present in WordNet.

	Input					
	50% terms			3 terms		
	P	R	F	P	R	F
English						
SingleLanguage	89.2	75.9	82.0	80.6	15.2	25.6
CrossLingual	86.5	91.1	88.7	86.7	60.2	71.1
Russian						
SingleLanguage	91.3	69.0	78.6	82.1	18.3	29.9
CrossLingual	84.9	86.2	85.5	85.3	62.1	71.9
Hebrew						
SingleLanguage	93.8	38.6	54.7	90.2	5.7	10.7
CrossLingual	86.5	82.4	84.4	93.9	55.6	69.8

Table 4: WordNet-based precision (P) and recall (R) for concept extension.

5.3 Contribution of each language

Each of the 45 languages we used influences the score of at least 5% of the discovered terms. However, it is not apparent if all languages are indeed beneficial or if only a handful of languages can be used. In order to check this point we have performed partial automated tests as described in Section 5.2, removing one language at a time. We also tried to remove random subsets of 2-3 languages, comparing them to removal of one of them. We saw that in each case removal of more languages caused a consistent (while sometimes minor) decrease both in precision and recall metrics. Thus, each language contributes to the system.

6 Discussion

We proposed a framework which given a set of terms defining a concept in some language, utilizes multilingual information available on the web in order to extend this list. This method allows to take advantage of web data in many languages, requiring only multilingual dictionaries. Our method was able to discover a substantially greater number of terms than state-of-the-art single language pattern-based concept extension methods, while retaining high precision.

We also showed that concepts obtained by this method tend to be more coherent in comparison to corresponding concepts in WN, a manually prepared resource. Due to its relative language-independence and modest data requirements, this framework allows gathering required

concept information from the web even if it is scattered among different and relatively uncommon or resource-poor languages.

References

Mishele Banko, Michael J Cafarella , Stephen Soderland, Matt Broadhead, Oren Etzioni, 2007. Open information extraction from the Web. *IJCAI '07*.

Wauter Bosma, Chris Callison-Burch, 2007. Paraphrase substitution for recognizing textual entailment.. *Evaluation of Multilingual and Multimodal Information Retrieval, Lecture Notes in Computer Science '07*.

Sharon Caraballo, 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. *ACL '99*.

Montse Cuadros, German Rigau, 2008. KnowNet: Building a large net of knowledge from the Web. *COLING '08*.

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, 2009. Translation and extension of concepts across languages. *EACL '09*.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, Richard Harshman, 1990. Indexing by latent semantic analysis. *J. 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 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 clusters. *EMNLP '04*.

Ilya Gelfenbeyn, Artem Goncharuk, Vladislav Lehelt, Anton Lipatov, Victor Shilo, 2003. Automatic translation of WordNet semantic network to Russian language (in Russian) *International Dialog 2003 Workshop*.

J. Gorman, J.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*.

Dekang Lin, 1998. Automatic retrieval and clustering of similar words. *COLING '98*.

Noam Ordan, Shuly Wintner, 2007. Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation 19(1):39-58, 2007*.

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*.

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*.

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*.

Rion Snow, Daniel Jurafsky, Andrew Ng, 2006. Semantic taxonomy induction from heterogeneous evidence. *COLING-ACL '06*.

Lonneke van der Plas, Jorg Tiedemann, 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. *COLING-ACL '06*.

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.*

Špela Vintar, Darja Fišer, 2008. Harvesting multiword expressions from parallel corpora. *LREC '08*.

Dominic Widdows, Beate Dorow, 2002. A graph model for unsupervised lexical acquisition. *COLING '02*.