Panlingual Lexical Translation via Probabilistic Inference

Mausam, Stephen Soderland, Oren Etzioni
Turing Center
Dept. of Computer Science and Engineering
University of Washington, Seattle, WA, 98195, USA
{mausam,soderlan,etzioni}@cs.washington.edu

Abstract

The bare minimum lexical resource required to translate between a pair of languages is a translation dictionary. Unfortunately, dictionaries exist only between a tiny fraction of the 49 million possible language-pairs making machine translation virtually impossible between most of the languages.

This paper summarizes the last four years of our research motivated by the vision of panlingual communication. Our research comprises three key steps. First, we compile over 630 freely available dictionaries over the Web and convert this data into a single representation – the translation graph. Second, we build several inference algorithms that infer translations between word pairs even when no dictionary lists them as translations. Finally, we run our inference procedure offline to construct PANDICTIONARY– a sense-distinguished, massively multilingual dictionary that has translations in more than 1000 languages. Our experiments assess the quality of this dictionary and find that we have 4 times as many translations at a high precision of 0.9 compared to the English Wiktionary, which is the lexical resource closest to PANDICTIONARY.

Introduction

Nearly 7,000 languages are in use today (Gordon 2005) out of which about 3,000 are endangered or even closer to extinction (Krauss 2007). With each dead language a whole cultural history is lost, a peek into an heritage of the bygone era is closed forever. Moreover, in the era of globalization, where inter-lingual communication is becoming increasingly important, one way the less-popular languages can survive is by having technology, particularly the machine translation (MT) systems, enable and facilitate this communication. Unfortunately, the current state of the art in MT, *e.g.*, Google Translate, which is able to handle on the order of only a thousand language pairs (out of 49 million), leaves a lot to be desired.

Because of its reliance on aligned corpora statistical MT is far from scaling the technology to this large number of language pairs. It is a pity, however, that the bare minimum of the lexical resources, a translation dictionary, is also not available between a large number of language pairs. This paper reports on our recent results in constructing PANDICTIONARY– a panlingual dictionary that can be used to translate words (or phrases) between any pair of languages (Etzioni *et al.* 2007; Mausam *et al.* 2009).

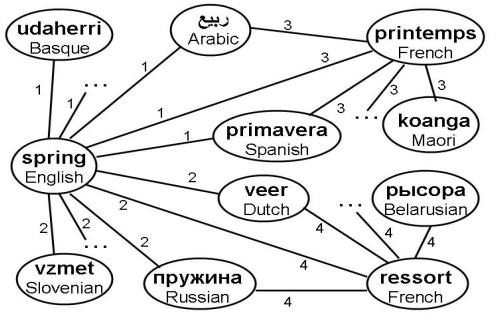

Figure 1: A fragment of the translation graph for two senses of the English word 'spring'. Edges labeled '1' and '3' are for spring in the sense of a season, and '2' and '4' are for the flexible coil sense. The graph shows translation entries from an English dictionary merged with ones from a French dictionary.

Of course, lexical translation cannot replace statistical MT, but it is useful for several applications including translating search-engine queries, meta-data tags in *flickr.com* and *del.icio.us*, library classifications and recent applications like cross-lingual image search (Etzioni *et al.* 2007) at *www.panimages.org*. Furthermore, lexical translation is a valuable component in knowledge-based Machine Translation (MT) systems, *e.g.*, (Carbonell *et al.* 2006) and is sufficient for lemmatic communication (Soderland *et al.* 2009).

This paper summarizes the following contributions:

1. We introduce a novel approach to the task of lexical translation, which compiles a large number of machine readable dictionaries in a single resource called a translation graph.
2. We employ probabilistic reasoning and inference over the translation graph to infer translations that are not expressed in any of the input dictionaries. We design several inference algorithms and compare their performance.
3. We use our best algorithm to compile PANDICTIONARY—a massive, sense-distinguished multilingual dictionary. Our empirical evaluations show that depending on the desired precision PANDICTIONARY is 4.5 to 24 times larger than the English Wiktionary (*http://en.wiktionary.org*). Moreover, it expresses about 4 times the number of pairwise translations compared to the union of its input dictionaries (at precision 0.8).

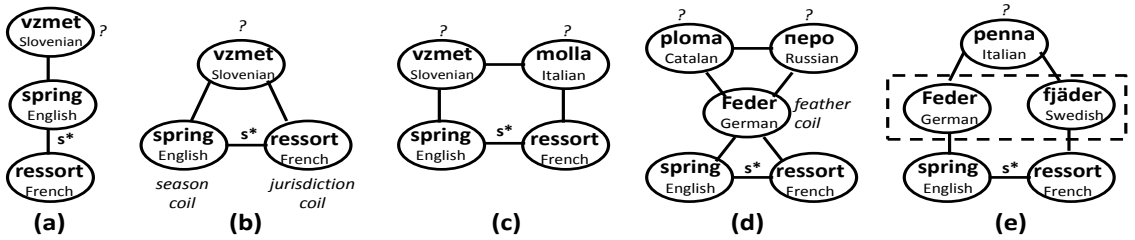

Figure 2: Snippets of translation graphs illustrating various inference scenarios. The nodes in question mark represent the nodes in focus for each illustration. For all cases we are trying to infer translations of the flexible coil sense of spring.

The Translation Graph

The translation graph is an undirected graph defined as $\langle \mathcal{V}, \mathcal{E} \rangle$. $\mathcal{V}$ and $\mathcal{E}$ denote the sets of vertices and edges. A vertex v represents a word or a phrase in a language. Edge between two vertices denotes that the two words have at least one word sense in common. Additionally, an edge is labeled by an integer denoting an ID for the word sense.

We build the translation graph incrementally on the basis of entries from multiple, independent dictionaries (both bi- and multi-lingual) hosted by the Web. Bilingual dictionaries translate words from one language to another, often without distinguishing the intended sense. The Wiktionaries (*wiktionary.org*) are multilingual dictionaries created by volunteers collaborating over the Web, which provide translations from a source language into multiple target languages, generally distinguishing between different word senses.

We assign each dictionary entry a unique sense ID. A sense-distinguished, multilingual entry is converted to a clique and all edges are assigned the same ID. As edges are added on the basis of entries from a new dictionary, some of the new word sense IDs are redundant because they are equivalent to word senses already in the graph from another dictionary. This leads to the following semantics for sense IDs: if two edges have the same ID then they represent the same sense, however, if two edges have different IDs, they may or may not represent the same sense.

Currently, our translation graph is compiled from more than 630 dictionaries, contains over 10,000,000 vertices and around 60,000,000 edges. It is truly panlingual – contains translations in over 1000 languages.

Figure 1 shows a fragment of a translation graph, which was constructed from two sets of translations for the word 'spring' from an English Wiktionary, and two corresponding entries from a French Wiktionary for 'printemps' (spring season) and 'ressort' (spring coil)[1]. Translations of the season 'spring' have edges labeled with sense ID=1, the coil sense has ID=2, translations of 'printemps' have ID=3, and so forth. Note that there are multiple IDs (1 and 3) that represent the season sense of 'spring' – we refer to this phenomenon as *sense ID inflation*.

Sense ID inflation poses a challenge for inference in translation graphs. If we wish to find all words that translate sense s^*, represented by a given ID, we need to look for evidence suggesting that another ID also represents s^*. We develop three algorithms for this task, which we describe next.

[1]Only a few edges are shown. For example, an edge between 'udaherri' and 'primavera' (ID 1) is present, but not shown.

Probabilistic Inference

Our inference task is defined as follows: given a sense ID, say id^*, that represents a sense, say s^*, compute the translations (in different languages) of s^*. We describe three algorithms for inference over the translation graph.

In essence, inference over a translation graph amounts to *transitive* sense matching: if word A translates to word B, which translates in turn to word C, what is the probability that C is a translation of A? If B is polysemous then C may not share a sense with A. For example, in Figure 2(a) if A is the French word 'ressort' (means both jurisdiction and the flexible-coil sense of spring) and B is the English word 'spring', then Slovenian word 'vzmet' may or may not be a correct translation of 'ressort' depending on whether the edge (B, C) denotes the flexible-coil sense of spring, the season sense, or another sense. However, if the three nodes form a triangle (Figure 2(b)) then our belief in the translation increases. This insight helps in our first inference algorithm.

TRANSGRAPH: In this method (Etzioni *et al.* 2007) we compute sense ID equivalence scores of the form $score(id_i \equiv id_j)$. The evidence to compute this equivalence comes from two sources: (1) if the vertex sets in two multilingual sense IDs have a high overlap the IDs are equivalent with a score proportional to the fraction of overlap, and (2) if two independent bilingual entries form a triangle with an edge labeled with id then two bilingual sense IDs are equivalent to id with a high score. Based on these sense ID equivalence scores each individual vertex can be scored – we follow a path from id^* to that vertex and multiply the sense ID equivalence scores at each hop. Ranking by this translation score gives us a way to trade precision for recall.

Theory of Translation Circuits: Continuing with the example of Figure 2 we question what is special about a triangle. In particular, can we make a similar inference in the snippet (c)? The answer is yes, under certain conditions detailed in (Mausam *et al.* 2009).

Definition 1 *We define a* translation circuit from v_1^* with sense s^* *as a cycle that starts and ends at v_1^* with no repeated vertices (other than v_1^* at end points). Moreover, the path includes an edge between v_1^* and another vertex v_2^* that also has sense s^* (examples are snippets (b) and (c)).*

Theorem 1 *Let C^k be a translation circuit of length k ($k << |\mathcal{S}|$) with origin v^* and sense s^*. Let P be the set of vertices along this circuit, let $|\mathcal{S}|$ denote the number of possible word senses for all words in all languages, and let the maximum number of senses per word be bounded by N ($N << |\mathcal{S}|$). Then under some as-*

sumptions $\forall v \in P \lim_{|\mathcal{S}|\to\infty} Pr(v \in s^*) = 1$

uSENSEUNIFORMPATHS: This theorem suggests the following basic algorithm: "for each vertex v check whether v lies on a translation circuit with sense s^* – if yes, mark it as a translation". In our algorithm we check for the presence of a translation circuit using a random walk scheme. Notice that this algorithm correctly infers 'Feder' to be translation of spring coil, and 'ploma' to be not (Figure 2(d)).

We additionally employ the observation that greater the number of translation circuits through a vertex greater our belief in the inference about it. To operationalize this we employ graph sampling – we sample different graph topologies by sampling each edge with a probability and check for the presence of translation circuit in each topology. If a vertex has circuits in many different sampled graphs it has more evidence and we assign a higher score for its inference.

SENSEUNIFORMPATHS: uSENSEUNIFORMPATHS achieves much more recall than TRANSGRAPH, but makes a specific kind of mistake. Figure 2(e) exemplifies this situation – our previous algorithm will incorrectly label 'penna' to be a translation of spring coil. Though it is not a translation, the circuit completes because two vertices 'Feder' and 'fjäder' have two senses in common – the spring coil sense and feather of a bird. 'Penna' means feather, but not coil. Shared polysemy in the circuits is the cause of many incorrect inferences.

Definition 2 *An* ambiguity set A *is a set of vertices that all share the same two senses. I.e.,* $\exists s_1, s_2,$ with $s_1 \neq s_2$ s.t. $\forall v \in A,\ v \in s_1 \wedge v \in s_2$.

In our example 'Feder' and 'fjäder' form an ambiguity set. To increase the precision of our algorithm we *prune* the circuits that contain two nodes in the same ambiguity set and also have one or more intervening nodes that are not in the ambiguity set. There is a strong likelihood that the intervening nodes will represent a translation error.

The key step that remains is "how to compute an ambiguity set"? Ambiguity sets can be detected from the graph topology automatically. Each clique in the graph represents a set of vertices that share a common word sense. When two cliques intersect in two or more vertices, the intersecting vertices share the word sense of both cliques. This may either mean that both cliques represent the same word sense, or that the intersecting vertices form an ambiguity set. A large overlap between two cliques makes the former case more likely; a small overlap makes it more likely that we have found an ambiguity set.

Experiments: Which of the three algorithms (TG, uSP and SP) is superior for translation inference? To carry out this comparison, we randomly sampled 1,000 senses from English Wiktionary and ran the three algorithms over them. We assess the precision and coverage of these inference algorithms by comparing the inferred translations with a gold standard. We create the gold standard on a subset of seven languages for which we had in-house experts.

Our results are shown in Figure 3. At this high precision, SP more than doubles the number of baseline translations, finding 5 times as many inferred translations (in black) as TG. The number of inferred translations (in black) for SP is

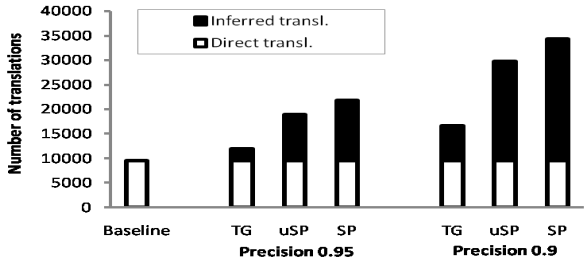

Figure 3: The SENSEUNIFORMPATHS algorithm (SP) more than doubles the number of translations at precision 0.95, compared to a baseline of translations that can be found without inference.

1.2 times that of uSP and 3.5 times that of TG, at precision 0.9. SP is consistently better than others, since it performs better for polysemous words, due to its pruning based on ambiguity sets. We conclude that SP is the best inference algorithm and employ it for further research.

PanDictionary: A Novel Multilingual Resource

To be most useful for our vision of panlingual translation we wish to construct a *sense-distinguished* lexical translation resource, in which each entry is a distinct word sense and associated with each word sense is a list of translations in multiple languages. This will enable lexical translation for a large number of languages at once just by looking up the desired sense. We compile PANDICTIONARY, a first version of such a dictionary, by employing SENSEUNIFORMPATHS over the translation graph.

We first run SENSEUNIFORMPATHS to expand the approximately 50,000 senses in the English Wiktionary. We further expand any senses from the other Wiktionaries that are not yet covered by PANDICTIONARY, and add these to PANDICTIONARY. This results in the creation of the world's largest multilingual, sense-distinguished translation resource, PANDICTIONARY. It contains a little over 80,000 senses. Its construction takes about three weeks on a 3.4 GHz processor with a 2 GB memory.

We evaluate PANDICTIONARY's quality and coverage across two dimensions. (1) We compare the coverage of PANDICTIONARY with the largest existing multilingual dictionary, the English Wiktionary? (2) We evaluate the benefit of inference over the mere aggregation of 631 dictionaries.

The English Wiktionary is the largest Wiktionary with a total of 403,413 translations.[2] It is also more reliable than some other Wiktionaries in making word sense distinctions. In the first study we use only the subset of PANDICTIONARY that was computed starting from the English Wiktionary senses. Thus, this experiment under-reports PANDICTIONARY's coverage.

To evaluate a huge resource such as PANDICTIONARY we recruited native speakers of 14 languages – Arabic, Bulgarian, Danish, Dutch, German, Hebrew, Hindi, Indonesian, Japanese, Korean, Spanish, Turkish, Urdu, and Vietnamese. We randomly sampled 200 translations per language, which resulted in about 2,500 tags. Figure 4 shows the no. of translations in PANDICTIONARY in senses from the English Wik-

[2]Our translation graph uses the version of English Wiktionary extracted in January 2008.

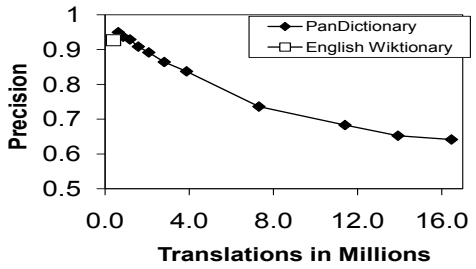

Figure 4: Precision *vs.* coverage curve for PANDICTIONARY. It quadruples the size of the English Wiktionary at precision 0.90, is more than 8 times larger at precision 0.85 and is almost 24 times the size at precision 0.7.

tionary. At precision 0.90, PANDICTIONARY has 1.8 million translations, 4.5 times as many as the English Wiktionary.

We also compare the coverage of PANDICTIONARY with that of the English Wiktionary in terms of languages covered. Table 1 reports, for each resource, the number of languages that have a minimum number of distinct words in the resource. PANDICTIONARY has 1.4 times as many languages with at least 1,000 translations at precision 0.90 and more than twice at precision 0.7. These observations reaffirm our faith in the panlingual nature of the resource.

Next, we investigate whether this increase in coverage is due to the inference algorithm or the mere aggregation of hundreds of translation dictionaries. Since most bilingual dictionaries are not sense-distinguished, we ignore the word senses and count the number of distinct (word1, word2) translation pairs. To create a gold standard for translations we use *collaborative tagging* scheme, with two native speakers of different languages, who are both bilingual in English. For each suggested translation they narrate in English the various senses of words in their respective languages. They tag a translation correct if they found a common sense, one that is shared by both the words.

Figure 5 compares the number of word-word translation pairs in the English Wiktionary (EW), in all 631 source dictionaries (631 D), and in PANDICTIONARY at precisions 0.90, 0.85, and 0.80. PANDICTIONARY increases the number of word-word translations by 73% over the source dictionary translations at precision 0.90 and increases it by 2.7 times at precision 0.85. PANDICTIONARY also adds value by identifying the word sense of the translation, which is not given in most of the source dictionaries.

Overall, our experiments demonstrate that PANDICTIONARY, which is our compiled dictionary, has much larger coverage than English Wiktionary, the largest multilingual dictionary known to us before this project. We also observe that our algorithms infer a large number of translations that are not in any of the input dictionaries quadrupling the number of pairwise translations asserted (at precision 0.8).

	# languages with distinct words		
	≥ 1000	≥ 100	≥ 1
English Wiktionary	49	107	505
PanDictionary (0.90)	67	146	608
PanDictionary (0.85)	75	175	794
PanDictionary (0.70)	107	607	1066

Table 1: PANDICTIONARY covers substantially more languages than the English Wiktionary.

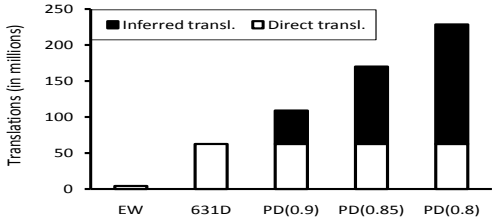

Figure 5: The number of distinct word-word translation pairs from PANDICTIONARY is several times higher than translation pairs in the English Wiktionary (EW) or in all 631 source dictionaries combined (631 D). A majority of PANDICTIONARY translations are inferred by combining entries from multiple dictionaries.

Related Work and Conclusions

Related Work: Because we are considering a relatively new problem (automatically building a panlingual translation resource) there is little work that is directly related to our own. Most previous work on compiling dictionaries automatically has focused on a relatively small set of languages (*e.g.*, (Helmreich, Guthrie, & Wilks 1993)). Previous algorithms for translation inference (Gollins & Sanderson 2001) are unable to achieve high precision or use additional sources of data like parallel corpora (Dyvik 2004).

Conclusions: We have described a novel approach to lexical translation that combines freely available dictionaries in a common resource, and runs probabilistic inference to infer new translations not mentioned in the source data. This leads to the construction of PANDICTIONARY, the largest multilingual sense-distinguished dictionary covering over 1000 languages. Our evaluation contrasts our high coverage with English Wiktionary, the closest multilingual dictionary in terms of size and scope. We plan to make PANDICTIONARY available to the research community, and also to the Wiktionary community to bolster their efforts.

Acknowledgments: This research was supported by a gift from Utilika Foundation to the Turing Center at Univ. of Washington.

References

Carbonell, J.; Klein, S.; Miller, D.; Steinbaum, M.; Grassiany, T.; and Frey, J. 2006. Context-based machine translation. In *AMTA*.

Dyvik, H. 2004. Translation as semantic mirrors: from parallel corpus to WordNet. *Language and Computers* 49(1):311–326.

Etzioni, O.; Reiter, K.; Soderland, S.; and Sammer, M. 2007. Lexical translation with application to image search on the Web. In *Machine Translation Summit XI*.

Gollins, T., and Sanderson, M. 2001. Improving cross language retrieval with triangulated translation. In *SIGIR*.

Gordon, Jr., R. G., ed. 2005. *Ethnologue: Languages of the World (Fifteenth Edition)*. SIL International.

Helmreich, S.; Guthrie, L.; and Wilks, Y. 1993. The use of machine readable dictionaries in the Pangloss project. In *AAAI Spring Symposium on Building Lexicons for Machine Translation*.

Krauss, M. E. 2007. Keynote-mass language extinction and documentation: The race against time. In *The Vanishing Languages of the Pacific Rim*. Oxford University Press.

Mausam; Soderland, S.; Etzioni, O.; Weld, D.; Skinner, M.; and Bilmes, J. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *ACL'09*.

Soderland, S.; Lim, C.; Mausam; Qin, B.; Etzioni, O.; and Pool, J. 2009. Lemmatic machine translation. In *MT Summit XII*.