One sense per collocation for prepositions

Hollis Easter & Benjamin Schak

May 7th, 2003

Abstract

This paper presents an application of the one-sense-per-collocation hypothesis to the problem of word sense disambiguation for prepositions. The hypothesis is tested through translation using a bilingual French-English corpus. The paper shows that one-sense-per-collocation does hold for prepositions.

1 Introduction

The one-sense-per-collocation hypothesis (Yarowsky 1993) states that words[1] tend to occur with only one sense within different instances of the same collocation. Yarowsky (1993) tested this hypothesis with strong results on coarse-grained senses of ambiguous nouns, verbs, an adjectives. Although Martinez and Agirre (2000) achieved weaker results for fine-grained sense distinctions, the hypothesis can help a wide range of natural language processing tasks. Since the one-sense-per-collocation hypothesis is implicit in much of the previous work, such as (Japkowicz, 1991) on translating prepositions, an evaluation of the Hypothesis could yield improvement in translation systems. This paper discusses compelling reasons for why the Hypothesis should hold, and tests the Hypothesis on a bilingual English-French corpus.

Our first problem is how to define senses for prepositions. Yarowsky (1993) gives several ways to approach this. One way is the "hand-tagged homograph method," in which one uses a corpus tagged with the correct senses of each word. This won't work for us because no corpus known to us has reliable sense distinctions for prepositions. We also want to avoid methods based on

[1] Words with more than one sense are polysemes.

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 8–14
Computer Science Department, Swarthmore College

homophones, ambiguities in online character recognition, and pseudo-words because the closed class of prepositions is too small. So, we equate the notion of a sense with that of a French translation.

1.1 Subcategorization

As noted above, there are two linguistic observations that recommend the one-sense-per-collocation hypothesis. The first of these is subcategorization, the notion that every noun, verb, and adjective selects (or "takes") certain types of phrases for complements, and can determine the heads of those complements. For example, consider the English adjective *interested*, translated into French as *interessé*. Sentences (1) and (2) show that *interested* must take a prepositional phrase headed by the preposition *in* as its complement, while *interessé* must take a prepositional phrase headed by *par*.

(1) John is interested *math / in math / *for math / *to math / *mathematic / *to do math.

(2) Jacques est interessé *les maths / par les maths / *pour les maths / * aux maths / *mathématique / *faire les maths.

It should be clear that there is nothing about mathematics *per se* that requires one preposition or another; while one can be interested *in* math, one can also rely *on* math or be afraid *of* math or look *to* math.

1.2 Noun-complement specificity

The second encouraging observation, used by Japkowicz and Wiebe (1991), is that many nouns may only be complements of certain prepositions. They assert that most nouns may only be used with particular prepositions, and that analogous nouns in different languages (English and French, for example) admit different propositions because the languages conceptualize those nouns differently. For example, in saying *on the bus* but *dans l'autobus* (literally "in the bus"), "English conceptualizes the bus as a *surface* that can *support* entities, by highlighting only its bottom platform, while French conceptualizes the bus as a *volume* that can *contain* entities, by highlighting its bottom surface, its sides, and its roof altogether." (Japkowicz, 1991)[2]

[2]Readers may wonder when prepositions are determined by a preceding word and when they are determined by a complement. We suspect that adverbial prepositional phrases, such as Jap-

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 8–14
Computer Science Department, Swarthmore College

1.3 Local collocations

In testing one-sense-per-collocation for nouns, verbs, and adjectives, Yarowsky (1993) tested only local collocations. That is, he ignored the possibility that distant content words could give reliable information sense disambiguation. We do the same here, and with better cause. While it is somewhat plausible that senses of nouns, verbs, and adjectives—categories whose words are replete with meaning—could be inferred from distant context, such a situation seems unlikely for prepositions.

1.4 Potential problems

Given these sensible arguments for the Hypothesis, why bother testing it? Trujillo (1992) provides examples where the one-sense-per-collocation hypothesis fails. He presents an English sentence (3) with three plausible Spanish translations (4).

(3) She ran under the bridge.

(4) Corrió debajo / por debajo / hasta debajo del puente.

The first translation implies that she was running around under the bridge, the second that she ran on a path that went under the bridge and kept going, and the third that she ran up to a position under the bridge and stopped. We hope, however, that this example is of an infrequent special case, and can be overcome. Sentence (3) usually translates best with *por debajo*, and the same sentence with the verb *rested* translates best with *debajo de*.

Another possible problem is that individual speakers may use different prepositional phrases for essentially the same concept. While one speaker may use *on top of*, another may use *atop*, another *on*, and so on. Given these issues, additional testing is warranted.

2 Methods

To test the Hypothesis, we used the sentence-aligned Hansards of the 36th Parliament of Canada, a French-English bilingual corpus. (*Hansards*, 2001) Our

kowicz and Wiebe's locatives, are determined by their complements, while prepositional phrases governed by a preceding noun, verb, or adjective are determined by their governor.

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 8–14
Computer Science Department, Swarthmore College

analysis takes four steps:

1. We preprocess the French sentences, changing *au* to *à le*, *aux* to *à les*, *du* to *de le*, *des* to *de les*, and *d'* to *de*.

2. We create a database, for each preposition in our list[3], with one record for each appearance in the training corpus (36.5 million words). Each record contains the surrounding four English words and the preposition's French translation.

3. We create a list, for each preposition, of English context words, along with the most frequent translation for the preposition given each context word.

4. We test our list's predictions on a held-out portion (4.5 million words) of the Hansard corpus. We also test the performance of a naïve translation algorithm for a baseline.

The first step is justified because in French a word like *au* is equivalent to the preposition *à* combined with the article *le*. Since combination with an article doesn't affect the sense of a preposition, this is fine to do.

In the second and fourth steps we need the correct translation of each English preposition. Since the Hansards are not word-aligned, this is difficult to do accurately. Consider the following sentence pair:

> I went to a library yesterday.
>
> Je suis allé à la bibliothèque hier.

We make the (rather large) assumption that if an English preposition is found n% of the way through a sentence, then its translation will be found n% of the way through its sentence as well. Since *to* is word number 2 (starting counting from 0) out of six words, and since the French sentence has seven words, our initial guess is that the translation of *to* is at position $2(7-1)/(6-1) \approx 2$. We find the word *allé* in that position, which is not an acceptable translation (taken from the *Collins-Robert French-English English-French Dictionary* (Atkins, 1996) of *to*. So, we look in the positions surrounding *allé*, and find *à*, an acceptable

[3]We use the prepositions *against*, *around*, *at*, *before*, *by*, *during*, *for*, *from*, *in*, *like*, *of*, *off*, *on*, *out*, *through*, *up*, and *with*. These were chosen because some are polysemous and some are monosemous, thereby providing a diverse set of test cases.

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 8–14
Computer Science Department, Swarthmore College

translation, and halt. (In fact, we give up after searching four words ahead and behind.) This approach seems to work fairly well for the Hansard corpus, in large part because of the stilted, literal translations in it. Clearly, a word-aligned corpus would make better predictions here, particularly in instances where either English or French uses a multi-word preposition (e.g., *off of* or *autour de*).

In the fourth step, we get a baseline by measuring how a naïve word-for-word translation does on our held-out corpus. We simply translate each English preposition with its most common (or at least most canonical) French translation: *at* to *à*, *in* to *dans*, and so on.

3 Results

We tabulated results for each preposition. The following are typical of our results:

for

Context	Precision	Accuracy
Two before	.9625	.6886
One before	.9564	.7027
One after	.9683	.6842
Two after	.8880	.6938
None	1.0000	.2857

of

Context	Precision	Accuracy
Two before	.9817	.9169
One before	.9795	.9175
One after	.9826	.9172
Two after	.8993	.9155
None	1.0000	.9181

The precision is the number of times our translation list made a prediction divided by the number of prepositions encountered in the testing corpus. The accuracy is the number of times our translation list made a *correct* prediction divided by the number of times it made any prediction. Clearly, the improvements are much greater for some prepositions than for others. The results for all prepositions combined are:

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 8–14
Computer Science Department, Swarthmore College

Total Context	Precision	Accuracy
Two before	.9457	.7936
One before	.9394	.8084
One after	.9510	.8190
Two after	.8618	.8166
None	1.0000	.6140

The results show that surrounding context includes sufficient information to improve translation of most prepositions into French. In general, context words closer to the preposition give better information. We find this somewhat strange, since the word directly after a preposition is often an article, which should contribute little sense information.

Different prepositions give much different results, as shown in the sample data above. Why, in particular, are our results for *of* so poor compared with the baseline? Suppose we are testing the $+1$ position for *of*. If the word after *of* in our testing corpus is *Parliament*, for example, our system will guess whatever the most common translation of *of* before *Parliament* was during training. Since *of* almost always translates as *de*, the guessed translation will be *de* for almost any context word, and therefore our accuracy results will be much like the baseline accuracy for *de*.

4 Conclusion

All four context positions (two before, one before, one after, and two after the English preposition) were helpful in translation, giving clear benefits over the baseline. However, the best results came from the word immediately after the preposition.

There are several ways to improve on these results. First, a word-aligned corpus would erase the error introduced by our translation-guessing algorithm. Second, we might improve results by looking at more than one context word at a time, or by weighting the predictions based on some context words higher than others. However, even our limited results show that the one-sense-per-collocation hypothesis is often reliable for English prepositions.

It is possible that idiomatic usage occurs in the Hansard corpus enough to throw off the results. Therefore, it would be interesting to see the preposition-

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 8–14
Computer Science Department, Swarthmore College

translation model applied to a number of different languages in parallel. At present, the lack of multilingual aligned corpora makes this infeasible, but should they become available, that experiment would have stronger results.

References

[Atk] Atkins, Beryl T., et al. *Collins-Robert French-English English-French Dictionary*, 3rd ed. New York: HarperCollins Publishers and Dictionnaires Le Robert, 1996.

[Han] Germann, Ulrich, ed. *Aligned* Hansards *of the 36th Parliament of Canada, Release 2001-1a*. 2001.

[Jap] Japkowicz, Nathalie, and Janyce Wiebe. "A System for Translating Locative Prepositions from English into French." *Meeting of the Association for Computational Linguistics* 1991: 153-160.

[Mart] Martinez, David, and Eneko Agirre. "One Sense per Collocation and Genre/Topic Variations." *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*: 207-215.

[Word] Princeton University, Cognitive Science Laboratory. *WordNet, version 1.7.1*.

[Tru] Trujillo, Arturo. "Spatial Lexicalization in the Translation of Prepositional Phrases." *30th Annual Meeting of the Association for Computational Linguistics* 1992: 306-308.

[Yar] Yarowsky, David. "One Sense per Collocation." *Proceedings, ARPA Workshop on Human Language Technology* 1993: 266-271.

Appeared in: *Proceedings of the Class of 2003 Senior Conference*, pages 8–14
Computer Science Department, Swarthmore College