Multilingual Corpus-based Approach to the Resolution of English –ing

Lee Schwartz (Microsoft Research, NLP)
Takako Aikawa (Microsoft Research, NLP)
One Microsoft Way, Redmond, WA 98052-6399, USA
(leesc@microsoft.com, takakoa@microsoft.com)

Abstract

Corpus data has proven to be useful for dealing with ambiguities in NLP. A number of studies, for example, have deal with disambiguating English PP attachments, using corpus data (Hindle and Rooth (1993), Brill and Resnik (1994), Steina and Nagao (1997), Ratnaparkhi (1998), and Pantel and Lin (2000), among others). This paper explores a novel approach to resolving ambiguities associated with *–ing* + Noun constructions in English. We use an aligned multilingual (English, Spanish, French, German and Japanese) corpus to extract lexical information necessary for disambiguation. Our premise is that while in English *–ing* constructions are highly ambiguous, corresponding constructions in other languages may not be ambiguous, and can thus provide English with disambiguating information. We argue that with aligned multilingual corpora, languages can learn non-trivial linguistic information from one another.

1. Ambiguities in English *–ing* constructions

Different syntactic and semantic relationships can exist in English between an *-ing* verb form and a following noun. At the syntactic level, an NLP system must decide whether the *–ing* + noun construction is a verb + object pair, or if it is a modifier + noun pair. So, for example, in (1a) *using* is a verb with the object *passwords*, whereas in (1b) *testing* is a modifier of *purposes*.

(1a) Click to learn more about *using passwords* with your identity.
(1b) For *testing purposes*, click Next.

For the purpose of translation, it is often the case that we need to specify what type of modification relationship exists between an *-ing* form and a following noun in a noun phrase. In (1b) the relationship of *testing* to *purposes* might be considered one of adjunct to noun as in the paraphrase, *purposes of testing*. But in other constructions that are similar with respect to syntax, the noun following the *-ing* form may actually be better thought of as the subject of the *-ing* verb. So, in (1c) the noun *rows* might be interpreted as the subject of *matching*, as in the paraphrase *rows that match*.

(1c) It specifies that *matching rows* returned by the query match a list of words.

Certainly, a similar paraphrase, i.e, *purposes that test*, is not possible for (1b). In this paper we explore the automatic extraction of information necessary to distinguish verb + object constructions (such as (1a)) from modifier + noun constructions (such as (1b) and (1c)).

2. *-Ing* constructions in other languages

While in English, the *-ing* + Noun construction is often ambiguous, in other languages, various linguistic devices, often unambiguous in nature, are used to instantiate the different relationships between the parts of the construction. For example, the NP *licensing information* in (2a), in which *licensing* is a modifier of the noun *information* (i.e., 'information for licensing'), is likely to be expressed as a compound noun in languages such as Japanese or German as shown in (2b) and (2c). In languages such as French or Spanish, on the other hand, the same type of modifier + noun relationship is likely to be expressed as a noun + prepositional phrase construction ('information about licensing'), as shown in (2d) and (2e).

(2a) English: When the number of users is different from the number of computers, this may provide incorrect *licensing information*.
(2b) Japanese: ユーザーの数がコンピュータの数と異なる場合は、正しい*ライセンス情報*が提供されない場合があります。
(2c) German: Wenn die Anzahl der Benutzer von der Anzahl der Computer abweicht, kann dies die *Lizenzinformationen* verfalschen.
(2d) French: Lorsque le nombre d'utilisateurs est different du nombre d'ordinateurs, cette procedure peut fournir *des informations* incorrectes *a propos des licences*.
(2e) Spanish: Cuando el numero de usuarios es distinto del numero de equipos, esto puede proporcionar *informacion de licencias* incorrecta.
A similar observation as to the clarity of the relationship between the parts of an *–ing* construction can be made for examples (3a)-(3e). The relationship between *entering* and the noun

name in (3a) is one of verb + object.[1] The verb + object relationship can be instantiated overtly in other languages: in Japanese/German, (3b)/(3c), the object noun is marked unambiguously as accusative. In French, (3d), because of the relative order of the noun *nom* and the participial verb *tapant*, *nom* is unambiguously the object of *tapant*. Finally, in Spanish, (3e), *nombre* is not identified unambiguously as the object of *escribir* (i.e., it could be the subject), but, at least, it is clear that the latter is a verb (and not a modifier of *nombre*).

(3a) English: You can rename the index by *entering a new name* in this box.

(3b) Japanese:インデックスの名前を変更したい場合は、このテキスト ボックスに*新しい名前を入力します*。

(3c) German: Sie konnen den Index umbenennen, indem Sie in diesem Feld *einen neuen Namen eingeben*.

(3d) French: Vous pouvez le renommer *en tapant un nouveau nom* dans cette zone.

(3e) Spanish: Para cambiar el nombre del indice, puede *escribir un nuevo nombre* en este cuadro.

3. Motivation

Determining the actual relationship between an *ing* form and a following noun based on a surface string is essential for understanding the string and for producing accurate translation into languages that do not have this ambiguity. Our work is conducted in the context of the MT project (MSR-MT) at Microsoft (Menezes & Richardson, 2001), which we describe below.

3.1. MSR-MT Overview

Our system consists of four main components: (i) Analysis (Parser); (ii) Logical Form (LF); (iii) Alignment/Transfer; and (iv) Surface String Generation. Our LF of the sentence in (4), for instance, is as follows:

(4) The purpose is stated in this file.

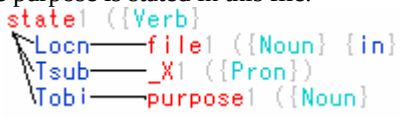

Figure 1

We use a large bilingual corpus of parsed sentences and align the LFs for each pair of sentences. The Alignment/Transfer phase of translation takes the aligned LF pairs and identifies correspondences between LF nodes. During translation, the system chooses the best translation mapping for the source sentence and transfers it into the target language system (Menezes & Richardson, 2001). The transferred LF, for example, produced for Japanese from the LF in Figure 1 is shown in Figure 2.

Transferred LF for Japanese

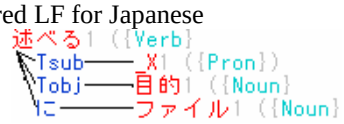

Figure 2

From this transferred LF, the generation component produces the sentence string in (5).

(5) 目的は、このファイルに述べられます。

3.2. Problem

Prior to this work, the English parser used in the MSR-MT system would produce either an Adjective + Noun analysis of an *--ing* + Noun construction, or a Verb + Noun analysis. For lack of information to the contrary, the parser could analyze the phrase *using digital signatures* in sentence (6) as an NP with *using* as an adjectival modifier of *digital signatures*. The LF for such a misanalysis would look like the one in Figure 3.[2]

(6) If your NAS supports *using digital signatures* for verification, click OK.

English LF of the misparsed sentence in (6)

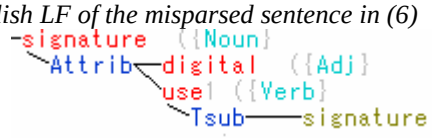

Figure 3

Here a misparse triggers an LF misanalysis, which, in turn, triggers a faulty Transferred LF as shown in Figure 4 for Japanese.

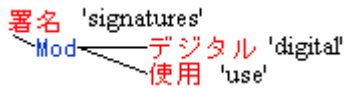

Figure4

The transferred LF incorrectly specifies a modifier (Mod) relationship between 署名 (*signatures*) and 使用 (*using*).

In the absence of information to the contrary, the parser could also misanalyze *ing* + noun constructions that are in a modifier + noun relationship as verb + object pairs. Such is the case below, in which *options* is misanalyzed as the object of *dial*.

[1] Note that this particular example is unambiguous in English because of the presence of the determiner with *name*. This would not be true, however, for *entering names*.

[2] An analysis of this type would be fine for the *matching rows* example in (1c).

(7) Click Dialing Options
(English LF of the misparsed sentence in (7))

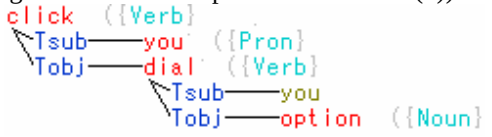

Figure 5

Such a misanalysis would lead, for example, to the Spanish translation in (8) as opposed to the correct translation in (9).

(8) Haga clic en *Marcar Opciones*.
('Click on Dial Options')

(9) Haga clic en *Opciones de marcado*.
('Click on Options of dialing')

4. Overview of Our Approach

We took an unsupervised approach to learning the relationships between an *–ing* word and a following noun. We used an aligned corpus from the computer domain which contains 74K sentences in five languages, Spanish, Japanese, French, German, and English.[3]

4.1. Methodology

We began by parsing all the sentences in the aligned multilingual corpus, obtaining LFs for them, and aligning these LFs on a pairwise basis, i.e, English-French (EF), English-Japanese (EJ), English-Spanish (ES), and English-German (EG). We then ran a filter on the pairwise aligned LFs to extract the portions of the LFs corresponding to the *ing* + Noun constructions. With the parser analyzing the *–ing* word of the construction as either an adjective or a verb, the LFs of the *–ing* + Noun construction had one of the following configurations:

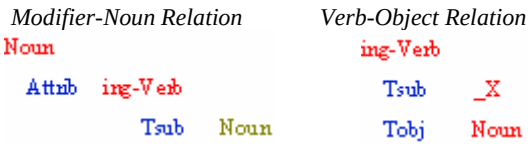

Figure 6

The filter then examined the relationship in the non-English LF between the word that was aligned with the English *ing* form and the word that was aligned with the English noun. For example, consider the English and French aligned sentences in (10) and their (aligned) LF fragments in Figure 7.

[3] Out of these 74 K sentences, we took out 14.8K sentences as test data and used the rest (59.2k) as training data.

(10) English: This may provide *licensing information*.
French: Cette procedure peut fournir *des informations a propos des licences*.

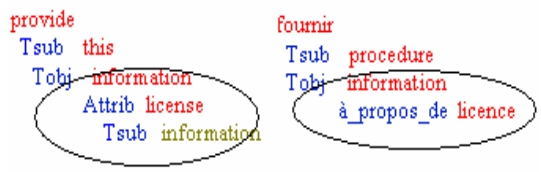

Figure 7

In the French LF, *information* dominates *licence* , and *licence* is not the subject of *information*. This is the basic condition for the filter's classification of *licensing information* as Modifier-Noun Relationship in Figure 6.

Note that this condition is also satisfied by the LF for the single German word, *Lizenzinformationen* and by the LF for the single Japanese word, ライセンス情報, both of which correspond to *licensing information*.

Figure 8

In these languages, such words are treated as compound nouns and hence, the part, *Lizenz* or ライセンス, which corresponds to the *-ing* word *licensing*, is represented as a Mod (Modifier) of *information* or 情報.

The basic condition, on the other hand, for the filter classifying an *-ing* construction as Verb-Object Relationship is that in the non-English LF, the word aligned to the *ing* form dominates the word aligned to the noun following the *ing* form, and the latter is not in a subject relationship to its parent.

4.2. *Ing* auxiliary dictionaries.

We collected sets of verb+object (TypObj) and adjunct+noun (NonObj) pairs from the aligned data from each of the language pairs (ES, EJ, EF, and EG) by running the filter described in the previous section over the aligned LFs for the sentences in the multilingual corpus. We then took the intersection of these sets as trusted data. From this trusted data we created two auxiliary, domain-specific dictionaries for testing purposes. The first dictionary included only the TypObj and NonObj pairs that were extracted from all four of the language pairs (4 lang-dictionary). The second

dictionary included the TypObj and NonObj pairs that occurred in at least three of the four language pairs (3 lang-dictionary).[4] To see how these dictionaries were built, consider, the following sampling of verb + object data extracted from our aligned bilingual texts.

(verb_object)	ES	EG	EF	EJ
connect(ing), device	no	yes	no	no
monitor(ing), method	no	yes	no	yes
access(ing), table	yes	yes	yes	no
create(ing), index	yes	yes	yes	yes

Table 1

The numbers of pairs extracted for each of the dictionaries are presented in Table 2. We made TypObj and NonObj information available to the parser via these auxiliary dictionaries.

	TypObj	NonObj
3 lang	1375	316
4 lang	339	155

Table 2

5. Evaluation

We measured the impact of the two *–ing* auxiliary dictionaries on our NLP system by evaluating the change in quality of translations from English into the four other languages. We used IBM BLEU (Papineni, et.al., 2002) to assess the quality of the MT translations, as BLEU scores seem to correlate with human assessments (Coughlin 2003). As test data, we used 2K English sentences from the technical domain. We translated these sentences once using an *–ing* auxiliary dictionary and once not using an *-ing* auxiliary dictionary. We then ran IBM BLEU on the two translations. Table 3 presents the results of the BLEU scores of these language pair translations.

	No dict	3 lang dict	4 lang dict
E-J	0.2319	0.2466	0.2454
E-S	0.3919	0.417	0.4164
E-G	0.1481	0.1745	0.1736
E-F	0.2426	0.2512	0.2512

Table 4

6. Concluding Remarks

We conclude that there is a significant difference between the quality of translations when English parsing is performed with the aid of an *ing* dictionary and when it is performed without the aid of an *ing* dictionary. We also conclude that the three-language dictionary produced better results than the four-language dictionary, i.e., it is not necessary that all four languages agree on an analysis for it to be trusted. We leave for future work the task of teasing apart various types of relationships underlying the modifier-noun constructions (e.g., *testing purposes* v.s. *singing birds*).

[4] The main reason for creating two dictionaries is that we wanted to allow for the fact that the expression of an *–ing* construction in a non-English language might actually be ambiguous.

References

Brill, E. and Resnik, P. (1994). A Rule-based Prepositional Phrase Attachment Disambiguation. In *Proceedings of COLING-94*, Kyoto, Japan.

Deborah Coughlin (2003). Correlating Automated and Human Assessments of Machine Translation Quality, In *Proceedings of MT Summit IV*.

Hindle, D and Rooth, M. (1993). Structural Ambiguity and Lexical Relations, *Computational Linguistics 19* (1): 103-120.

Menezes, A. and Richardson, S. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the ACL 2001*, Toulouse, France.

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, pp. 311-318. Philadelphia, PA.

Pantel, P and Lin, D. (2000). An Unsupervised Approach to Prepositional Phrase Attachment using Contextually Similar Words. In *Proceedings of Association for Computational Linguistics 2000*, pp. 101-108, Hong Kong.

Ratnaparkhi, A. (1998). Unsupervised Statistical Models for Prepositional Phrase Attachment. In *Proceedings of COLING-ACL 98*, Montreal, Canada.

Steina, J. and Nagao, M. (1997). Corpus Based PP Attachment Ambiguity Resolution with a Semantic Dictionary. In *Proceedings of the Fifth Workshop on Very Large Corpora*, pp. 66-80, Beijing and Hong Kong.