Enriching Word Alignment with Linguistic Tags

Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie M. Strassel, Kazuaki Maeda
xuansong@ldc.upenn.edu niyuge@us.ibm.com sgrimes@ldc.upenn.edu strassel@ldc.upenn.edu maeda@ldc.upenn.edu

Abstract

To improve automatic word alignment and ultimately machine translation quality, an annotation framework is jointly proposed by LDC and IBM. The framework enriches word alignment corpora to capture contextual, syntactic and language-specific features by introducing linguistic tags to the alignment annotation which indicates minimum translation units and translation relations of parallel text. A large amount of Chinese-English data was manually word-aligned and tagged following this framework. Evaluation of joint pilot annotation shows high inter-annotator agreement rate of above 90%. [1]

1 Introduction

In machine translation, word alignment is a crucial intermediate stage indicating corresponding word relations in parallel text. Traditionally, statistical word alignment models are unsupervised algorithms (Brown et al. 1993, Melamed 2000). These models rely on a considerable amount of data to learn coherent language phenomena. More recently, with the availability of manually word-aligned data, supervised methods such as the Maximum Entropy based models (Ittercheriah et. al. 2005) have shown promising results. Supervised algorithms typically employ linguistic features such as part-of-speech and parse information. Empirical results show that MaxEnt models outperform traditional models in word alignment quality. Motivated by such improvement, Linguistic Data Consortium (LDC) collaborated with IBM in a pilot study to design and streamline a unified framework for linguistically-enriched word alignment annotation corpora. This paper describes the motivation and the details of the framework, and is organized in the following way. Sections 2 and 3 detail alignment and tagging methodologies of the framework. Section 4 focuses on Chinese-English corpora. Section 5 presents an evaluation of joint pilot annotation between IBM and LDC. Section 6 concludes with future work.

2 Alignment Methodology

Our alignment framework establishes rules for alignment annotation which is further enriched with linguistic tags from the tagging framework. Two approaches were previously proposed for word alignment: *minimum match* and *attachment*. In our framework, these two methods are further refined and more precisely and consistently applied. The refinement allows us to achieve higher annotation agreement rates and it fits more tightly in the new tagging framework.

The goal of *minimum match* is to find complete and minimal semantic translation units. These minimal translation units are atomic translation pairs and cannot be further decomposed into sub-part links. In Chinese-English alignment, the minimal atomic translation unit is one-to-one link built on one character. However, there are frequent cases where the characters are inseparable from each other and must be made into a single unit. Abbreviations, idiomatic expressions, set or frozen expressions are a few examples. In the past, the minimum match approach did not apply to such cases. Instead, each case was specially treated by ad hoc rules. We overcome this shortcoming by consistently applying the minimum approach to many-to-one and many-to-many links, generating minimal linguistic units in addition to one-character links.

The attachment approach is adopted for handling unaligned words. Translating cross-cultural thought inevitably involves translation adaptations

[1] This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

and variations which change surface structures. Words added/omitted in the surface structure, as a result, have no matching equivalents in alignment. Deleting them, however, corrupts the correctness of the sentence. They are contextually or functionally required for semantic equivalence. Studying these omissions/additions unveils special grammatical and translation rules. Previously, the attachment rules apply only to function words and were very loosely defined. There were no specific rules or linguistic tests to help the annotators decide whether a function word should be attached to its head or left unaligned. In our new framework, attachment rules are more rigorously defined, removing much of the annotation ambiguity. Attachment thus can apply to both function and content words if their equivalents are not present in the translation. Phrase-level extra words are attached to their constituent head words to indicate phrasal constituent dependency or collocation dependency. For sentence-level extra words, attachment approach indicates sentence syntactic dependency relations. Grouped phrasal constituents and attached words are further tagged using link and word tags according to different functions they assume, which further helps with disambiguation.

3 Tagging Methodology

The goal of tagging is to alleviate word insertion and deletion problems in statistical translation models by providing linguistic information on alignment links and unaligned words. We designed 8 link tags and 14 word tags to systematically address a variety of linguistic phenomena, including context-free lexical, context-dependent, syntactic, and language specific features.

3.1 Tagging Links

3.1.1 Context-free Link Tags

There are two tags for context-free links, *semantic* or *function.* They facilitate the extraction of context-free lexical translation pairs which can be readily re-used in machine translation systems and other natural language processing applications. The interpretation of these context-free links involves no or minimal contextual clues because they are atomic and cannot be further decomposed into sub-part links. Semantic links look for semantic word similarity while function links show function resemblance. Semantic links refer to links between content words. A link is *semantic* if both sides are content words. Otherwise, it is a *function* link. The aligned pairs may be one-to-one, one-to-many, many-to-one, or many-to-many. The multi-character unit pairs such as idioms or set/frozen expressions are also context-free semantic link. The minimum approach is employed for finding such atomic translation pairs.

3.1.2 Composite Link Tags

In contrast to context free atomic translation pair, unaligned words are attached to their constituent heads to form composite links. We distinguish two types of composite links: *grammatically-inferred* and *contextually-inferred* links. Similar to in-context translation (ICT) (Ker and Chang, 1997) pairs, contextually-inferred links capture contextual equivalence. The interpretation of context-dependent features/clues has special value for supervised approaches in enhancing translation quality. Explicitly incorporating contextual features such as translation association clues (Tiedemann, 2003) significantly reduces alignment error rate. Grammatically-inferred links align functional or grammatical equivalence. Revealing word relations, these links also capture syntactic constituent dependency features of the source and target languages. Incorporation of dependency features into phrase-based models promotes machine translation quality (Och and Ney, 2002). Such supervised models rely heavily on hand-aligned bilingual corpora for syntactic constituent relations.

3.1.3 Specific Link Tags

Traditional statistical translation models are language-independent and usually fail to tackle problems occurred due to language specific features. Finding alignment relations between parallel texts becomes all the more challenging due to language idiosyncrasies. Capturing idiosyncratic features help machine translation learn better models. Chinese “的”, for instance, is notoriously hard to deal with. In our tagging framework, we tag all instances of “的”. First, to capture *clause marker, possessive* and *preposition* functions, three “的” link tags are designed: DE-modifier link, DE-

clause link, and DE-possessive link. Other usages of "的" are captured by grammatically-inferred links where unaligned "的" assumes functions like tense, passive voice, modification, possessive or statement function, which is further tagged with word tags. Several other Chinese specific features, such as tense, measure words or 把-structure, are also annotated using grammatical-inferred link tags.

3.2 Tagging Words

While tagging aligned links aims to map symmetric deep-structural semantic equivalence, tagging unaligned words inside links describes asymmetric surface-structure divergences which contribute to such semantic equivalence. The extra attached words can be function or content words, providing grammatical or contextual/semantic clues.

3.2.1 Tagging Content Words

Extra attached words inside *contextual-inferred* links are usually content words, without which, the structure might be grammatical, but it is not semantically sensible. They are *obligatory* for the context. To build up semantic/contextual equivalence, they are required because of word association rules or collocation conventions of a particular language. "*Local context marker*" is a word tag to indicate this feature. It is applied at the local phrase or sentence level. For the unaligned words at a sentence/discourse level, we use "context obligatory" and "context non-obligatory" word tags to show if they are contextually required or not.

3.2.2 Tagging Function Words

Compared to unaligned content words, unaligned function words occur more frequently in translation, which is a very difficult problem for machine translation. To better describe syntactic features, we designed 10 word tags to handle unaligned and attached function words for Chinese-English tagging. The tags include *tense/passive marker, omni-function-preposition marker, DE-modifier marker, possessive marker, to-infinitive marker, sentence marker, measure word marker, determiner/demonstrative marker, clause marker, anaphoric reference marker, and rhetorical marker.* Some of these tags can be universal such as the tense tag. Others are language specific, revealing idiosyncratic features of Chinese and English. For instance, the functions of unaligned "的" are captured by tense/passive marker, DE-modifier markers, possessive marker, and sentence marker. The particular subject-drop feature in Chinese is captured by anaphoric reference marker.

4 Chinese-English Word-aligned and Tagged Corpora

Following this unified alignment and tagging framework, we have created large amount of manually word-aligned and tagged corpora enriched with linguistic tags. This is an on-going project at LDC.

Language	Genre	Files	CharTokens	Segments
Chinese	Newswire	579	225645	5015
Chinese	Broadcast conversation	10	86356	3328
Chinese	Weblog	360	191098	8211
Total			503099	16554

Figure 1. Annotation Data Profile

Raw data need to be segmented for alignment. In this framework, the Chinese word segmentation is done on the smallest linguistic unit. In case of Chinese, that unit is "character". This is one of the simplest kinds of word segmentation, each character being a word. In most machine translation systems of Chinese-English, more sophisticated word segmentation schemes are used to group characters into "words". We distinguish these two types of segmentation by denoting the first type *character segmentation* and the latter *word segmentation*. One of the benefits of aligning at character level is to enable machine translation systems to define source language 'words' (e.g. Chinese). One way to do this is to define Chinese word as a sequence of contiguously aligned characters to the same English word. Another benefit is that character-level word alignments can easily support other higher-level larger component alignments. The tagging task is based on this character level alignment.

5 Evaluation

We measure annotation agreement by computing precision, recall, and F-score. Figure 2 shows the inter-annotator agreement of joint annotation between LDC and IBM on alignment, and Figure 3

on tagging annotation inside LDC. The intra-annotator agreement is 95%-98%.

Number of links	File1(101) A1(106) A2(105)	File2(174) A1(182) A2(182)	File3(238) A1(262) A2(257)	File4(197) A1(220) A2(216)
Precision	95%	95%	91%	90%
Recall	96%	95%	93%	91%
F-score	96%	95%	92%	91%

Figure 2. Inter-AA on Alignment

Number of tags	1078 tags File1	421 tags File2	SemanticLink A1(112) A2(115)	FuncLink A1(66) A2(56)
A1 & A2	94%	93%	n/a	n/a
Precision	n/a	n/a	97%	85%
Recall	n/a	n/a	95%	100%
F-score	n/a	n/a	96%	92%

Figure 3. Inter-AA on Tagging

6 Future Work

Future task will scale to more systematic classification of multi-level linguistic tags. Recently, this framework has also been successfully applied to Arabic-English word alignment task with coarse tags. In the future, richer tags will be defined. We also want to explore the portability of this framework to other language pairs other than Arabic and Chinese. In addition, larger higher-level constituent component alignment will be automated by post-processing character-level alignment.

As the linguistic resources described above are distributed to GALE program participants, LDC will wherever possible distribute the data broadly through usual mechanisms. The tagging specification is available on LDC's website (http://projects.ldc.upenn.edu/gale/task_specifications/). The annotated corpora will be made generally available as regular LDC publications over time.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistcal machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

I. D. Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

A. Ittycheriah, S. Roukos, 2005. A Maximum Entropy Word Aligner for Arabic-English Machine Translation, In *Proceedings of HLT/EMNLP*, pages 89–96, Vancouver.

S.J. Ker, J.S. Chang. 1997. A Class-base Approach to Word Alignment, Computational Linguistics, Vol. 23, No. 2, pp. 313-343.

J. Tiedemann. 2003. Combining clues for word alignment. In *Proceedings of EACL*.

Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th ACL*, pages 295-302.