# Enriching Word Alignment with Linguistic Tags

**Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie M. Strassel, Kazuaki Maeda**

Linguistic Data Consortium, University of Pennsylvania
Philadelphia, PA 19104 USA
IBM T.J. Watson Research Center
Yorktown Heights, N.Y. 10598 USA
Email: {xuansong,sgrimes,strassel,maeda}@ldc.upenn.edu niyuge@us.ibm.com

**Abstract**

Incorporating linguistic knowledge into word alignment is becoming increasingly important for current approaches in statistical machine translation research. To improve automatic word alignment and ultimately machine translation quality, an annotation framework is jointly proposed by LDC (Linguistic Data Consortium) and IBM. The framework enriches word alignment corpora to capture contextual, syntactic and language-specific features by introducing linguistic tags to the alignment annotation. Two annotation schemes constitute the framework: alignment and tagging. The alignment scheme aims to identify minimum translation units and translation relations by using minimum-match and attachment annotation approaches. A set of word tags and alignment link tags are designed in the tagging scheme to describe these translation units and relations. The framework produces a solid ground-level alignment base upon which larger translation unit alignment can be automatically induced. To test the soundness of this work, evaluation is performed on a pilot annotation, resulting in inter- and intra- annotator agreement of above 90%. To date LDC has produced manual word alignment and tagging on 32,823 Chinese-English sentences following this framework.

## 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 & Roukos, 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, LDC collaborate 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 the joint pilot annotation by 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 links built on one character as shown in Figure 1 where each character is aligned one English word.


Figure 1: One-to-One Links

However, there are frequent cases where the characters are inseparable from each other and must be made into a single unit, which is also a minimal atomic translation unit. Abbreviations, idiomatic expressions, set or frozen expressions are a few examples. Figure 2 is an example of many-to-many alignment.

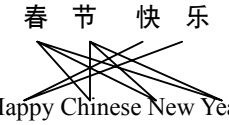


Figure 2: Many-to-Many Links

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 and variations which change surface structures. As a result, words added/omitted in the surface structure, have no matching equivalents in alignment. Deleting them, however, would corrupt the correctness of the sentence. They are contextually or functionally required for semantic equivalence. Consider the examples in Figure 3.

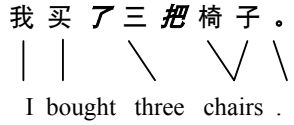


Figure 3a: Unaligned Chinese

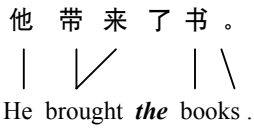


Figure 3b: Unaligned English

In 3a and 3b, the Chinese measure word "把" and "了" cannot be aligned to any English word, and in 3b "the" in the English does not have its Chinese equivalent. These words are "extra". They are inserted or added to make the sentences grammatically acceptable or contextually complete.

Studying these omissions/additions unveils special grammatical and translation rules. Previously, the attachment rules applied 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. Therefore, the "把" in 3a is attached to the Chinese "three" because Chinese measure words usually co-occur with numbers, and the English "the" is attached to "books" to show its constituent head. "了" is attached to the main verbs to contextually match the past tense words "bought " and "brought" in the English sentences. With these specific attachment rules, annotators would more consistently handle the unaligned words at the local or phrase level. These attachments are shown in Figure 3c and 3d.

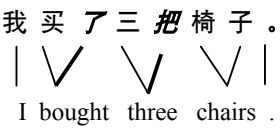


Figure 3c: Attachment of Unaligned Chinese

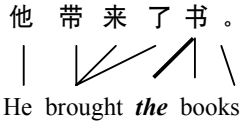


Figure 3d: Attachment of Unaligned English

Extra words at the sentence or discourse level, on the other hand, have no immediate constituents to rely on or attach to. However, they are still needed grammatically or contextually. For such words, rather than applying the attachment approach, we directly tag these words to indicate if they are grammatically or contextually needed.

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 eight link tags and fourteen 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. A link is *semantic* if both sides are content words. Otherwise, it is a *function* link. Semantic links look for semantic word similarity while function links show function resemblance. Links such as 'chairs' in Figure 3a are semantic links. Figure 4 is an example of function link.

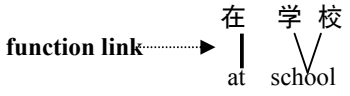


Figure 4: Function Link

Semantic links refer to links between content words. 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 links. The minimum approach is employed for finding such atomic translation pairs. Figure 5 shows an example of tagging of many-to-many semantic links.


Figure 5: Many-to-many Semantic Link

**3.1.2 Composite Link Tags**

In contrast to context-free atomic translation pairs, composite links are formed by attaching unaligned words to their constituent heads. All attached words in the links are further tagged with appropriate word tags.

We distinguish two types of composite links: *grammatically-inferred* and *contextually-inferred* links. Grammatically-inferred links apply to those words that have no translation counterpart but are required to make the sentence grammatical. The alignment links of these words are called grammatically-inferred links. In Figure 6, the grammatically-inferred Chinese "将" has no English translation but is necessary to make the Chinese sentence grammatical. In addition to link tags, the word "将" also gets a word tag. (§3.2.2) Figure 6 does not show all links for clarity.

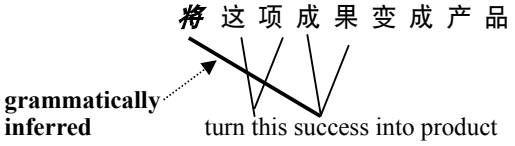


Figure 6: Grammatically-inferred Links

While function words are often involved in grammatically-inferred links, *content* words are also occasionally omitted from or inserted to surface structure because of word association/collocation (or pragmatic) feature. They are normally inferred via collocation or association context. Without them, the sentence may be grammatically or structurally acceptable, but not semantically sensible. We call such alignment "*contextually-inferred* links". In Figure 7, "收看" (meaning "watch") is a content word inferred from the association of "CCTV". The attached content words are further tagged with a word tag. (§3.2.1)


Figure 7: Contextually-inferred Links

In-context translation (ICT) (Ker & Chang, 1997) and the interpretation of context-dependent features have 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. Composite links in our work align functional, grammatical and contextual equivalence. In addition to 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 improves machine translation quality (Och & Ney, 2002). Such supervised models will benefit the most from our new alignment framework because they rely heavily on hand-aligned bilingual corpora for syntactic constituent relations.

**3.1.3 Language 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. Language specific features can be easily defined and implemented in our tagging framework. In this section, we use Chinese as a working example.

Chinese "的" is a notoriously hard word to deal with because of the wide range of linguistic functions "的" assumes. In our tagging framework, we define tags for each of these functions. In this way, we are able to tag all instances of "的". When "的" appears in an alignment link, the link is of one of the three types shown in Table 1, which are illustrated with examples in Figure 8.

| 的 (DE) function | Link Type |
|---|---|
| In a relative clause | DE-clause |
| In a prepositional phrase | DE-modifier |
| In a possessive construct | DE-possessive |

Table 1: Chinese-specific Tags

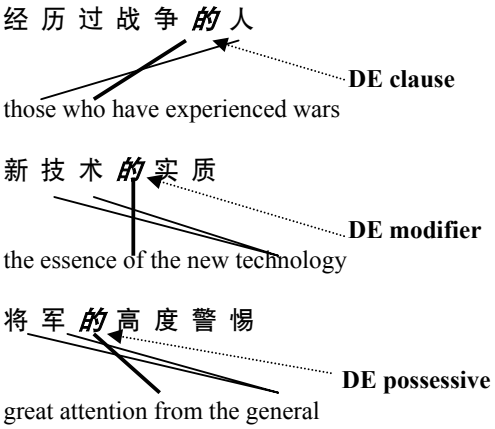


Figure 8: Chinese-specific Tags

### 3.2 Tagging Words

While tagging links aims to map symmetric deep-structural semantic equivalence, tagging 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

Attached words inside *contextual-inferred* links are usually content words, without which, the structure might be grammatical, but it is not semantically acceptable or complete. Figure 9 shows such an example.

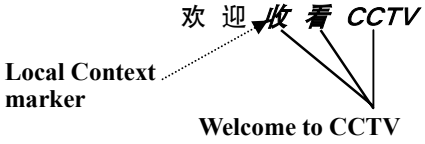


Figure 9: Local Context Markers

In Figure 9, "收看" is *obligatory* for the context. It is required in Chinese in this context, without which the meaning would be different. "*Local context marker*" is a word tag to indicate this feature. It is applied at the local phrase level and such words usually have local constituent words to attach to or rely on.

#### 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 design eleven word tags to handle attached function words for Chinese-English tagging. The tags along with examples are shown in Tables 2 and 3. Table 2 shows general-purpose tags. In Table 2, highlighted words in the examples are tagged with the corresponding tag. English examples are shown for those tags that are more prominent in English and vice versa for Chinese. Chinese translations are shown in square brackets. Table 3 shows tags specific to Chinese "的".

| Word Tag | Example |
|---|---|
| Tense/Passive marker | He **is** eating. |
| Omni-Function-Preposition | … **被**(by) 别人(someone) 粉 碎 (shatter) [someone shattered…] |
| DE-modifier marker | 干 (do) **得** 好 (well) [did well] |
| Possessive marker | the head **of** the branch |
| TO-infinitive marker | continue **to** work |
| Sentence marker | … 什么(what) **呢**? [ what is …?] |
| Measure Word marker | **一 根** (one) 柱 子 (pillar) [one pillar] |
| Determiner/Demonstrative marker | **The** main purpose |
| Clause marker | The mistake **which** he made |
| Anaphoric Reference marker | President Clinton said that **he** would … |
| Rhetorical marker | 大 陆 (mainland) **专 家** (expert) 和 (and) 台 湾 (Taiwan) **专 家** (expert) [experts from mainland and Taiwan] |

Table 2: General-purpose Tags

Some of the tags are designed to capture universal language pattern or features such as the tense tag while others reveal more idiosyncratic language features such as DE("的")-modifiers.

| Word Tag | Example |
|---|---|
| Tense/Passive marker | 提 交 (submit) **的** 报 告 (report) [report submitted] |
| DE-modifier marker | 红(red)红(red)**的** 。[red] |
| Possessive marker | 宾(Penn)大(university)**的** 教 授 (professor) [Penn professor] |
| Sentence marker | 卫生员(medics)是(are)相 当 (quite) 忙 (busy) **的** 。(medics are quite busy.) |

Table 3: Chinese-的 Tags

#### 3.2.3 Tagging Words at Discourse Level

*Context-obligatory* and Non-c*ontext-obligatory* tags are used for unaligned words at a sentence/discourse level to indicate whether or not they are contextually required. Context obligatory words are grammatically and/or semantically inferred words. They are needed but they have no local constituent unit to rely on or attach to. We use "context obligatory maker" for these unaligned words at the sentence or discourse level. In Figure 10, "it" and "you do" are obligatory in English, indicating the feature of subject drop in Chinese. All such words are tagged with "context-obligatory" word tag.

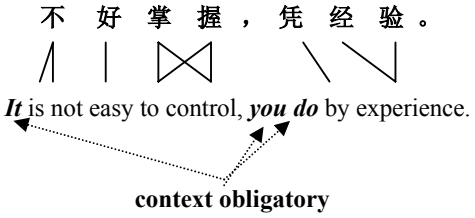


Figure 10: Context Obligatory Markers

*Non-context-obligatory* word tags are used for unaligned words which are neither grammatically required nor

semantically needed. They are extra words used for smoothing the tone, without which, the sentences are still grammatical and the meaning remains the same. For instance, the "也" and "都" in Figure 11 are not contextually obligatory and they can be removed without affecting meaning and grammar.


Figure 11: Non-context-obligatory Markers

## 4. Chinese-English Word-aligned and Tagged Corpora

### 4.1 Data Profile

Using this unified alignment and tagging framework, we have created a large amount of manually word-aligned Chinese-English corpora enriched with linguistic tags. This is an on-going project at LDC. Table 4 illustrates the amount of data annotated to date. "CharToken" represents total count of Chinese characters. "Segments" represents the total count of sentences.

| Genre | File | CharToken | Segment |
|---|---|---|---|
| Newswire | 579 | 225645 | 5015 |
| Broadcast News | 28 | 183400 | 6376 |
| Broadcast Conversation | 34 | 306497 | 12050 |
| Weblog | 747 | 229799 | 9382 |
| Total | 1388 | 945341 | 32823 |

Table 4: Annotation Data Profile

### 4.2 Segmentation

Raw data need to be segmented for alignment. In this framework, the 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.

### 4.3 Using the Data

The word aligned corpora built following this alignment-tagging framework can be flexibly used in multiple language processing applications to serve different needs of users. For instance, if the user is interested in extracting translation lexicons, the context-free links will be the focus. If the user would like to capture the syntactic and the contextual information for their advanced translation models, the composite links would be of special value. If users are concerned that their current model is not yet ready to digest the syntactic- and contextual- rich information, they can choose to decompose the composite links by automatically moving the attached words outside the alignment links, and all the links would become context-free links. This is possible because we tag all the attached words inside composite links. Figure 12 shows how the alignments of 'has' and '了' can be removed based on their tags.

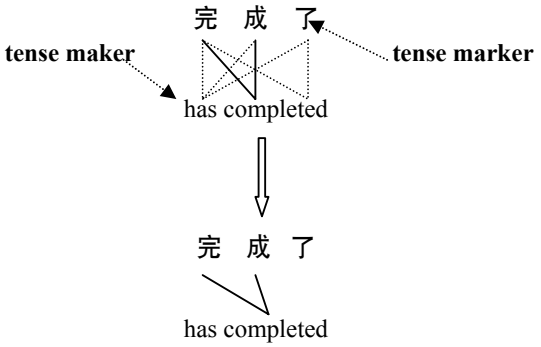


Figure 12: Removing Links Based on Tags

## 5. Evaluation

We measure annotation quality by computing precision, recall, and F-score. *Precision* is a measure of exactness, and *recall* measures completeness. The F-score, combining *precision* and *recall,* is the harmonic mean of the two as in Equation 1.

$$F = \frac{2 * precision * recall}{precision + recall} \quad (1)$$

Table 5 shows the inter-annotator agreement of the pilot annotation on alignment. Four newswire files are selected for annotation by two annotators, A1 and A2. Annotations by A1 serves as reference against which A2's annotations are measured. *Precision* is defined as the number of common links divided by the total number of A2's links (Common Links/A2). *Recall* is defined as the

number of common links divided by the total number of A1's links (Common Links/A1). The figures in Table 5 and Table 6 represent respectively the agreement of first round annotation by two junior annotators and the agreement of second round annotation by two senior annotators. After the second round quality-check annotation, the agreement scores are all above 90%. When the translation quality is not good enough, the increase in scores is not significant after the second round of annotation. In NW3 and NW4, there are longer sentences and the translation quality is not as good as that of NW1 and NW2.

| Data Source | Character Count | Precision | Recall | F-score |
|---|---|---|---|---|
| NW1 | 306 | 85.26% | 87.57% | 86.39% |
| NW2 | 185 | 87.50% | 86.73% | 87.11% |
| NW3 | 365 | 85.12% | 83.94% | 84.30% |
| NW4 | 431 | 81.68% | 82.90% | 82.29% |

Table 5: Inter-Annotator Agreement of First-round Alignment

| Data Source | Character Count | Precision | Recall | F-score |
|---|---|---|---|---|
| NW1 | 306 | 97.27% | 95.70% | 96.48% |
| NW2 | 185 | 95.28% | 96.19% | 95.73% |
| NW3 | 365 | 90.37% | 91.20% | 90.78% |
| NW4 | 431 | 90.83% | 92.61% | 91.17% |

Table 6: Inter-Annotator Agreement of Second-round Alignment

To test the inter-annotator agreement on tagging words and link types, we select two of the above aligned newswire files (NW1 and NW2) for other two junior annotators to do a round of tagging annotation. Table 7 shows the agreement on tagging all the Chinese characters, English words and alignment pairs. "Identical Tag" indicates that for a given word, character or aligned link, the two annotators either agree to assign identical tags or agree on the judgment to assign no tag.

| Data Source | Chi. Char Count | Eng. Word Count | Aligned Link Count | Identical Tag | Agreement |
|---|---|---|---|---|---|
| NW1 | 306 | 233 | 186 | 683 | 94.21% |
| NW2 | 185 | 131 | 105 | 392 | 93.11% |

Table 7: Inter-Annotator Agreement on Tagging

We choose to measure the agreement on semantic and function links to see how well the two annotators agree on a particular type of link. From the above tagged NW1 and NW2, we extract the semantic and function links separately and then compare if they are assigned to the same alignment. Table 8 shows that the agreement on function links is lower than that of the semantic links.

| Link Type | Precision | Recall | F-score |
|---|---|---|---|
| Semantic | 97.32% | 94.78% | 96.03% |
| Function | 84.62% | 98.21% | 90.90% |

Table 8: Inter-Annotator Consistency on Tagging Links

Table 9 shows the intra-annotator agreement of alignment with intervals of one week, two weeks, and one month.

| Interval | Precision | Recall | F-score |
|---|---|---|---|
| 1 week | 97.90% | 98.41% | 98.15% |
| 2 weeks | 96.31% | 97.86% | 97.07% |
| 1 month | 95.26% | 93.78% | 94.51% |

Table 9: Intra-annotator Agreement

## 6. Future Work

Future task will scale to more systematic classification of 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 and applied to Arabic-English word alignment to capture syntactic, contextual and Arabic-specific features. We also want to explore the portability of this framework to other language pairs other than Arabic and Chinese. Another possible direction is the higher-level constituent component alignment automated by post-processing character-level alignment. The approaches proposed here for word-level alignment may also be applied to larger component alignment to capture aligning and contextual features of higher levels, such as the phrase or/and tree levels.

As the linguistic resources described above are distributed to GALE (Global Autonomous Language Exploitation) program participants, LDC will wherever possible distribute the data more broadly, for example to its members and licensees, through the usual mechanisms. The alignment and tagging specifications and other details of the annotation approach are already available on the website of LDC (http://projects.ldc.upenn.edu/gale/task_specifications/) while the annotated corpora will be made generally available as regular LDC publications over time.

## 7. Acknowledgements

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.

## 8. References


Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, *19*(2), pp. 263-311.

Ittycheriah, A. & S. Roukos. (2005). A Maximum Entropy World Aligner for Arabic-English Machine Translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 89-96.

Ker, S. J. & Chang, J. S. (1997). A Class-base Approach to World Alignment. *Computational Linguistics*, *23*(2), pp. 313-343.

Li, X., Ge, N., Strassel, S.M., (2009). Guidelines for Chinese-English Word Alignment. Linguistic Data Consortium, University of Pennsylvania

Li, X., Ge, N., Strassel, S.M., (2009). Tagging Guidelines for Chinese-English Word Alignment. Linguistic Data Consortium, University of Pennsylvania

Melamed, I.D. (2000). Models of Translational Equivalence Among Words. *Computational Linguistics*, 26(2), pp. 221-249

Och, F. J., & Ney, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 295-302.

Tiedemann, J. (2003). Combining Clues for Word Alignment. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, *1*, pp. 339-346.