Aligning words in English-Hindi parallel corpora

Niraj Aswani
Department of Computer Science
University of Sheffield
Regent Court 211, Portobello Street
Sheffield S1 4DP, UK
N.Aswani@dcs.shef.ac.uk

Robert Gaizauskas
Department of Computer Science
University of Sheffield
Regent Court 211, Portobello Street
Sheffield S1 4DP, UK
R.Gaizauskas@dcs.shef.ac.uk

Abstract

In this paper, we describe a word alignment algorithm for English-Hindi parallel data. The system was developed to participate in the shared task on word alignment for languages with scarce resources at the ACL 2005 workshop, on "Building and using parallel texts: data driven machine translation and beyond". Our word alignment algorithm is based on a hybrid method which performs local word grouping on Hindi sentences and uses other methods such as dictionary lookup, transliteration similarity, expected English words and nearest aligned neighbours. We trained our system on the training data provided to obtain a list of named entities and cognates and to collect rules for local word grouping in Hindi sentences. The system scored 77.03% precision and 60.68% recall on the shared task unseen test data.

1 Introduction

This paper describes a word alignment system developed as a part of shared task on word alignment for languages with scarce resources at the ACL 2005 workshop on "building and using parallel texts: data driven machine translation and beyond". Participants in the shared task were provided with common sets of training data, consisting of English-Inuktitut, Romanian-English, and English-Hindi parallel texts and the participating teams could choose to evaluate their system on one, two, or all three language pairs. Our system is for aligning English-Hindi parallel data at the word level. The word-alignment algorithm described here is based on a hybrid – multi-feature approach, which groups Hindi words locally within a Hindi sentence and uses dictionary lookup (DL) as the main method of aligning words along with other methods such as Transliteration Similarity (TS), Expected English Words (EEW) and Nearest Aligned Neighbors (NAN). We used the training data supplied to derive rules for local word grouping in Hindi sentences and to find Named Entities (NE) and cognates using our TS approach. In the following sections we briefly describe our approach.

2 Training Data

The training data set was composed of approximately 3441 English-Hindi parallel sentence pairs drawn from the EMILLE (Enabling Minority Language Engineering) corpus (Baker et al., 2004). The data was pre-tokenized. For the English data, a token was a sequence of characters that matches any of the "Dr.", "Mr.", "Hon.", "Mrs.", "Ms.", "etc.", "i.e.", "e.g.", "[a-zA-Z0-9]+", words ending with apostrophe and all special characters except the currency symbols £ and $. Similarly for the Hindi, a token consisted of a sequence of characters with spaces on both ends and all special characters except the currency symbols £ and $.

3 Word Alignment

Given a pair of parallel sentences, the task of word alignment can be described as finding one-to-one, one-to-many, and many-to-many correspondences

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 115–118,
Ann Arbor, June 2005.

between the words of source and target sentences. It becomes more complicated when aligning phrases of one language with the corresponding words or phrases in the target language. For some words, it is also possible not to find any translation in the target language. Such words are aligned to null.

The algorithm presented in this paper, is a blend of various methods. We categorize words of a Hindi sentence into one of four different categories and use different techniques to deal with each of them. These categories include: 1) NEs and cognates 2) Hindi words for which it is possible to predict their corresponding English words 3) Hindi words that match certain pre-specified regular expression patterns specified in a rule file (explained in section 3.3.) and finally 4) words which do not fit in any of the above categories. In the following sections we explain different methods to deal with words from each of these categories.

3.1 Named Entities and Cognates

According to WWW1, the Named Entity Task is the process of annotating expressions in the text that are "unique identifiers" of entities (e.g. Organization, Person, Location etc.). For example: "Mr. Niraj Aswani", "United Kingdom", and "Microsoft" are examples of NEs. In most text processing systems, this task is achieved by using local pattern-matching techniques e.g. a word that is in upper initial orthography or a Title followed by the two adjacent words that are in upper initial or in all upper case. We use a Hindi gazetteer list that contains a large set of NEs. This gazetteer list is distributed as a part of Hindi Gazetteer processing resource in GATE (Maynard et al., 2003). The Gazetteer list contains various NEs including person names, locations, organizations etc. It also contains other entities such as time units – months, dates, and number expressions. Cognates can be defined as two words having a common etymology and thus are similar or identical. In most cases they are pronounced in a similar way or with a minor change. For example "Bungalow" in English is derived from the word "बंगला" in Hindi, which means a house in the Bengali style (WWW2). We use our TS method to locate such words. Section 3.2 describes the TS approach.

3.2 Transliteration Similarity

For the English-Hindi alphabets, it is possible to come up with a table consisting of correspondences between the letters of the two alphabets. This table is generated based on the various sounds that each letter can produce. For example a letter "c" can be mapped to two letters in Hindi, "क" and "स". This mapping is not restricted to one-to-one but also includes many-to-many correspondences. It is also possible to map a sequence of two or more characters to a single character or to a sequence two or more characters. For example "tio" and "sh" in English correspond to the character "श" in Hindi.

Prior to executing our word alignment algorithm, we use the TS approach to build a table of NEs and cognates. We consider one pair of parallel sentences at a time and for each word in a Hindi sentence, we generate different English words using our TS table. We found that before comparing words of two languages, it is more accurate to eliminate vowels from the words except those that appear at the start of words. We use a dynamic programming algorithm called "edit-distance" to measure the similarity between these words (WWW3). We calculate the similarity measure for each word in a Hindi sentence by comparing it with each and every word of an English sentence. We come up with an m x n matrix, where m and n refer to the number of words in Hindi and English respectively. This matrix contains a similarity measure for each word in a Hindi sentence corresponding to each word in a parallel English sentence. From our experiments of comparing more than 100 NE and cognate pairs, we found that the word pairs should be considered valid matches only if the similarity is greater than 75%. Therefore, we consider only those pairs which have the highest similarity among the other pairs with similarity greater than 75%. The following example shows how TS is used to compare a pair of English-Hindi words. For example consider a pair "aswani → अासवानी" and the TS table entries as shown below:

A→अ, S→स, SS→स, V→व, W→व and N→न

We remove vowels from both words: "aswn → असवन", and then convert the Hindi word into possible English words. This gives four different combinations: "asvn", "assvn", "aswn" and "asswn". These words are then compared with the actual English word "aswn". Since we are able to locate at least one word with similarity greater than 75%, we consider "aswani → आसवानी" as a NE. Once a list of NEs and cognates is ready, we switch to our next step: local word grouping, where all words in Hindi sentences, either those available in the gazetteer list or in the list derived using TS approach, are aligned using TS approach.

3.3 Local Word Grouping

Hindi is a partially free order language (i.e. the order of the words in a Hindi sentence is not fixed but the order of words in a group/phrase is fixed). Unlike English where the verbs are used in different inflected forms to indicate different tenses, Hindi uses one or two extra words after the verb to indicate the tense. Therefore, if the English verb is not in its base form, it needs to be aligned with one or more words in a parallel Hindi sentence. Sometimes a phrase is aligned with another phrase. For example "customer benefits" aligns with "ग्राहक के फायदे". In this example the first word "customer" aligns with the first word "ग्राहक" and the second word "benefits" aligns with the third word "फायदे". Considering "customer satisfaction" and "ग्राहक के फायदे" as phrases to be aligned with each other, "के" is the word that indicates the relation between the two words "ग्राहक" and "फायदे", which means the "benefits of customer" in English. These words in a phrase need to be grouped together in order to align them correctly. In the case of certain prepositions, pronouns and auxiliaries, it is possible to predict the respective Hindi postpositions, pronouns and other words. We derived a set of more than 250 rules to group such patterns by consulting the provided training data and other grammar resources such as Bal Anand (2001). The rule file contains the following information for each rule:

1) Hindi Regular Expression for a word or phrase. This must match one or more words in the Hindi sentence.
2) Group name or a part-of-speech category.
3) Expected English word(s) that this Hindi word group may align to.
4) In case a group of one or more English words aligns with a group of one or more Hindi words, information about the key words in both groups. Key words must match each other in order to align English-Hindi groups.
5) A rule to convert Hindi word into its base form.

We list some of the derived rules below:

1) Group a sequence of [X + Postposition], where X can be any category in the above list except postposition or verb. For example: "For X" = "X के लिये", where "For" = "के लिये".
2) Root Verb + (रहा, रही or रहे) + (PH). Present continuous tense. We use "PH" as an abbreviation to refer to the present/past tense conjunction of the verb "होना" - हुं, हैं, है, हो, etc.
3) Group two words that are identical to each other. For example: "अलग अलग", which means "different" in English. Such bi-grams are common in Hindi and are used to stress the importance of a word/activity in a sentence.

Once the words are grouped in a Hindi sentence, we identify those word groups which do not fit in any of the TS and EEW categories. Such words are then aligned using the DL approach.

3.3 Dictionary lookup

Since the most dictionaries contain verbs in their base forms, we use a morphological analyzer to convert verbs in their base forms. The English-Hindi dictionary is obtained from (WWW4). The dictionary returns, on average, two to four Hindi words referring to a particular English word. The formula for finding the lemma of any Hindi verb is: infinitive = root verb + "ना". Since in most cases, our dictionary contains Hindi verbs in their infinitive forms, prior to comparing the word with the unaligned words, we remove the word "ना" from the end of it. Due to minor spelling mistakes it is also possible that the word returned from dictionary does not match with any of the words in

a Hindi sentence. In this case, we use edit-distance algorithm to obtain similarity between the two words. If the similarity is greater than 75%, we consider them similar. We use EEW approach for the words which remain unaligned after the DL approach.

3.4 Expected English words

Candidates for the EEW approach are the Hindi word groups (HWG) that are created by our Hindi local word grouping algorithm (explained in section 3.3). The HWGs such as postpositions, number expressions, month-units, day-units etc. are aligned using the EEW approach. For example, for the Hind word "बावन" in a Hindi sentence, which means "fifty two" in English, the algorithm tries to locate "fifty two" in its parallel English sentence and aligns them if found. For the remaining unaligned Hindi words we use the NAN approach.

3.5 Nearest Aligned Neighbors

In certain cases, words in English-Hindi phrases follow a similar order. The NAN approach works on this principle and aligns one or more words with one of the English words. Considering one HWG at a time, we find the nearest Hindi word that is already aligned with one or more English word(s). Aligning a phrase "customer benefits" with "ग्राहक के फायदे" (example explained in section 3.3) is an example of NAN approach. Similarly consider a phrase "tougher controls", where for its equivalent Hindi phrase "अधिक नियंत्रण", the dictionary returns a correct pair "controls → नियंत्रण", but fails to locate "tougher → अधिक". For aligning the word "tougher", NAN searches for the nearest aligned word, which, in this case, is "controls". Since the word "controls" is already aligned with the word "नियंत्रण", the NAN method aligns the word "tougher" with the nearest unaligned word "अधिक".

4 Test Data results

We executed our algorithm on the test data consisting of 90 English-Hindi sentence pairs. We obtained the following results for non-null alignment pairs.

Word Alignment Evaluation
Evaluation of SURE alignments Precision = 0.7703 Recall = 0.6068 F-measure = 0.6788
Evaluation of PROBABLE alignments Precision = 0.7703 Recall = 0.6068 F-measure = 0.6788
AER = 0.3212

References

Bal Anand, 2001, *Hindi Grammar Books for standard 5 to standard 10*, Navneet Press, India.

Baker P., Bontcheva K., Cunningham H., Gaizauskas R., Hamza O., Hardie A., Jayaram B.D., Leisher M., McEnery A.M., Maynard D., Tablan V., Ursu C., Xiao Z., 2004, *Corpus linguistics and South Asian languages: Corpus creation and tool development,* Literary and Linguistic Computing, 19(4), pp. 509-524.

Maynard D., Tablan V., Bontcheva K., Cunningham H., 2003, *Rapid customisation of an Information Extraction system for surprise languages*, ACM Transactions on Asian Language Information Processing, Special issue on Rapid Development of Language Capabilities: The Surprise Languages.

WWW1, Named Entity Task Definition, http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_2.html#HEADING1 [15/04/2005]

WWW2, Britannica Online Encyclopaedia, http://www.britannica.com/eb/article?tocId=9018081 [15/04/2005]

WWW3, Dynamic Programming Algorithm (DPA) for Edit-Distance, http://www.csse.monash.edu.au/~lloyd/tildeAlgDS/Dynamic/Edit/ [22/03/05]

WWW4, English-Hindi dictionary source, http://sanskrit.gde.to/hindi/dict/eng-hin_guj.itx [22/03/05].