A statistical machine translation (MT) system may include a compound splitting module to split compounded words for more accurate translation. The compound splitting module selects a best split for translation by the MT system.

20 Claims, 5 Drawing Sheets
US 7,711,545 B2

Page 2

U.S. PATENT DOCUMENTS

5,870,706 A 2/1999 Alshawi
5,903,858 A 5/1999 Saraki
5,987,404 A 11/1999 Della Pietra et al.
5,991,710 A 11/1999 Papineni et al.
6,031,984 A 2/2000 Walser
6,032,111 A 2/2000 Mohri
6,082,034 A 7/2000 McCarley et al.
6,119,077 A 9/2000 Shozinok
6,131,082 A 10/2000 Hargrave et al.
6,182,014 B1 1/2001 Kenyon et al.
6,182,027 B1 1/2001 Nasukawa et al. ................................ 704/2
6,205,456 B1 3/2001 Nakaok
6,223,150 B1 4/2001 Duan et al.
6,236,958 B1 5/2001 Lange et al.
6,278,967 B1 8/2001 Akers et al.
6,289,302 B1 9/2001 Kuo
6,304,841 B1 10/2001 Berger et al.
6,311,152 B1 10/2001 Bai et al.
6,382,287 B1 5/2002 Poznanski et al.
6,393,388 B1 5/2002 Franz et al.
6,393,389 B1 5/2002 Chao et al.
6,415,250 B1 7/2002 van den Akker
6,460,151 B1 10/2002 Heidergton et al.
6,502,064 B1 1/2002 Miyahira et al.
6,782,356 B1 8/2004 Lopko
6,810,374 B2 10/2004 Kang

FOREIGN PATENT DOCUMENTS

EP 0933712 A2 8/1999
JP 07244666 1/1995
JP 100111148 1/1998
JP 11272672 10/1999

OTHER PUBLICATIONS

Al-Onaizan et al., “Translating with Scarce Resources,” 2000, 17th National Conference of the American Association for Artificial Intelligence, Austin, TX, pp. 672-678.


Yamamoto et al., “A Comparative Study on Translation Units for Bilingual Lexicon Extraction,” 2001, Japan Academic Association for Copyright Clearance, Tokyo, Japan.


* cited by examiner
Generate split options for German sentence g

Rank split based on frequencies of its parts in corpus

Check if splitting option(s) have translation in the English translation of the sentence

Use secondary translation lexicon to account for special cases

Qualify split based on statistics of parts-of-speech in the corpus

Select best split
FIG. 4

Aktionsplan

Aktion s plan

Aktion s plan

Aktion s plan
FIG. 5

break into known German words

find correspondences in English translation with help from translation lexicon
EMPIRICAL METHODS FOR SPLITTING COMPOUND WORDS WITH APPLICATION TO MACHINE TRANSLATION

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority to U.S. Provisional Application Ser. No. 60/484,812, filed on Jul. 2, 2003, the disclosure of which is incorporated here by reference in its entirety.

ORIGIN OF INVENTION

The research and development described in this application were supported by DARPA under grant number N66001-00-1-8914. The U.S. Government may have certain rights in the claimed inventions.

BACKGROUND

Machine translation (MT) is the automatic translation from a first language (a “source” language) into another language (a “target” language). Systems that perform an MT process are said to “decode” the source language into the target language.

A statistical MT system that translates foreign language sentences, e.g., French, into English may include the following components: a language model that assigns a probability P(e) to any English string; a translation model that assigns a probability P(θ|e) to any pair of English and French strings; and a decoder. The decoder may take a previously unseen sentence e and try to find the e that maximizes P(θ|e), or equivalently maximizes P(e|P(θ|e)).

Compounded words may present a challenge for MT systems. Compounding of words is common in a number of languages (e.g., German, Dutch, Finnish, and Greek). An example of a compounded word is the German word “Aktionplan”, which was created by joining the words “Aktion” and “Plan”. Words may be joined freely in such languages, which may greatly increase the vocabulary size of such languages.

SUMMARY

A statistical machine translation (MT) system may include a compound splitting module to split compounded words (“compounds”) for more accurate translation. The compound splitting module select a best split for translation by the MT system.

The compound splitting module may identify split option (s) for a compound, rank the compounds, and then pick a best translation from the compound and split option(s). The compound splitting module may rank using different metrics, e.g., frequency of a split’s parts in a corpus or translations of the compound in a translation lexicon. The compound splitting module may exclude split options based on parts-of-speech they contain, e.g., prepositions and determiners.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a statistical machine translation (MT) system including a compound splitting module according to an embodiment.

FIG. 2 is a block diagram of a compound splitting module.

FIG. 3 is a flowchart describing a compound splitting operation.

FIG. 4 shows splitting options for the compounded German word “Aktionplan”.

FIG. 5 shows correspondences between the splitting options for “Aktionplan” and the English translation.

DETAILED DESCRIPTION

FIG. 1 illustrates a statistical machine translation (MT) system 100 according to an embodiment. The MT system 100 may be used to translate from a source language (e.g., French) to a target language (e.g., English). The MT system 100 may include a language model 105, a translation model 110, and a decoder 115.

The MT system 100 may be based on a source-channel model. The language model 105 (or “source”) may assign a probability P(e) to any given English sentence e. The language model 105 may be an n-gram model trained by a large monolingual corpus to determine the probability of a word sequence. The translation model 110 may be used to determine the probability of correctness for a translation, e.g., the probability P(θ|e) of a French string θ given an English string e. The parameter values for computing P(θ|e) may be learned from a parallel corpus including bilingual sentence pairs. The translation model 110 may be, for example, an IBM translation Model 4, described in U.S. Pat. No. 5,477,451. The decoder may be used to identify the best translation by maximizing the product of P(e|P(θ|e)).

Compounding of words is common in a number of languages (e.g., German, Dutch, Finnish, and Greek). The compounded words (or “compounds”) may greatly increase the vocabulary size of such languages, which may present a challenge for MT systems.

In an embodiment, the MT system 100 may include a compound splitting module 120 to determine if and how a compound word should be split in a translation operation. FIG. 2 shows various components of the compound splitting module 120. These components may include a split generator 205, a frequency module 210, a primary translation lexicon 215, a secondary translation lexicon 220, a part-of-speech (POS) module 225, and a split selector 230.

FIG. 3 is a flowchart describing operations that may be performed by the compound splitting module in an MT system for translating German sentences into English. The split generator 205 may split a German word into possible split options (or “splits”) (block 305), e.g., into parts that have individual translations into English words. The frequency module 210 may select split(s) based on the frequencies of the splits’ parts in the corpus (block 310). The primary translation lexicon 215 may check if the splits have corresponding translations in the English translation of the sentence (block 315), and the secondary translation lexicon 320 may be used to account for special cases (block 320). The POS module 325 may qualify the splits based on statistics of parts-of-speech in the translation lexicon (block 325). The split selector 230 may then select the best split (block 330).

The split generator 205 may use known words, e.g., words existing in a training corpus 150 (FIG. 1) to identify possible splittings of a compound. In an experiment, the training corpus used was Europarl, which is derived from the European parliament proceedings and consists of 20 million words of German (available at http://www.isi.edu/publications/europarl/). To speed up word matching, the known words may be stored in a hash table based on the first three letters. The known words in the hash table may be limited to words having at least three letters.

The split generator 205 may account for filler letters between words in the compound. For example, the letter “s” is a filler letter in “Aktionplan”, which is a compound of the words “Aktion” and “Plan”. The filler letters “s” and “es” may
be allowed when splitting German words, which covers most cases. The splits may be generated using an exhaustive recursive search. As shown in FIG. 4, the split generator may generate the following splits for “Aktionsplan”: “aktions-plan”, “aktion-plan”, “aktions-plan”, and “akt-ion-plan”. Each part of the splits (i.e., “aktionsplan”, “aktions”, “aktion”, “akt”, “ion”, and “plan”) exist as whole words in the training corpus.

The frequency module 210 may identify the split having a highest probability based on word frequency. Given the count of words in the corpus, the frequency module may select the split S with the highest geometric mean of word frequencies of its parts $p_i$ ($n$ being the number of parts):

$$\text{arg max } \left( \prod_{i=1}^{n} \text{count}(p_i) \right)^{1/n}$$

The frequency module 210 utilizes a metric based on word frequency. The metric is based on the assumption that the more frequent a word occurs in a training corpus, the larger the statistical basis to estimate translation probabilities, and the more likely the correct translation probability distribution will be learned. However, since this metric is defined purely in terms of German word frequencies, there is not necessarily a relationship between the selected option and correspondence to English words. If a compound occurs more frequently in the text than its parts, this metric would leave the compound unbroken, even if it is translated in parts into English. In fact, this is the case for the example “Aktionsplan”. As shown in Table 1, the mean score for the unbroken compound (852) is higher than the preferred choice (825.6).

<table>
<thead>
<tr>
<th>Frequency of parts</th>
<th>Mean score</th>
</tr>
</thead>
<tbody>
<tr>
<td>aktionsplan (852)</td>
<td>852</td>
</tr>
<tr>
<td>aktion (960), plan (710)</td>
<td>825.6</td>
</tr>
<tr>
<td>aktion (960), plan (710)</td>
<td>825.6</td>
</tr>
<tr>
<td>aktion (224), ion (1), plan (710)</td>
<td>54.2</td>
</tr>
</tbody>
</table>

On the other hand, a word that has a simple one-to-one correspondence to English may be broken into parts that bear little relation to its meaning. For example, the German word “Freitag” (English: “Friday”) may be broken into “frei” (English: “free”) and “Tag” (English: “day”), as shown in Table 2.

<table>
<thead>
<tr>
<th>Frequency of parts</th>
<th>Mean score</th>
</tr>
</thead>
<tbody>
<tr>
<td>frei (885), tag (1864)</td>
<td>1284.4</td>
</tr>
<tr>
<td>freitag (556)</td>
<td>556</td>
</tr>
</tbody>
</table>

The translation lexicons may be used to improve one-to-one correspondence with English. The primary translation lexicon 215 may check for each split whether that split’s parts have translations in the English translation of the foreign language sentence(s) in the parallel corpus containing the compound. In the case of “Aktionsplan”, the words “action” and “plan” would be expected on the English side, as shown in FIG. 5. In case of “Freitag” the words “free” and “day” would not be expected. This information may be used by the compound splitting module 120 to break up “Aktionsplan”, but not “Freitag”.

The primary translation lexicon 215 may be learned from the parallel corpus 150. This can be done with the toolkit Giza, which establishes word-alignments for the sentences in the two languages. The toolkit Giza is described in Al-Omran et al., “Statistical machine translation,” Technical report, John Hopkins University Summer Workshop (1999).

To deal with noise in the translation table, the primary translation lexicon 215 may require that the translation probability of the English word given the German word be at least 0.01. Also, each English word may be considered only once. If a word is taken as evidence for correspondence to the first part of the compound, that word is excluded as evidence for the other parts. If multiple options match the English, the one(s) with the most splits may be selected and word frequencies may be used as a tie-breaker.

While this method has been found to work well for the examples “Aktionsplan” and “Freitag”, it failed in an experiment for words such as “Grundrechte” (English: “basic rights”). This word should be broken into the two parts “Grund” and “Rechte”. However, “Grund” translates usually as “reason” or “foundation”. But here, the more correct translation is the adjective “basic” or “fundamental”. Such a translation only occurs when “Grund” is used as the first part of a compound.

The second translation lexicon 220 may be used to account for such special cases. German words in the parallel corpus 150 may be broken up with the frequency method. Then, the translation lexicon may be trained using Giza from the parallel corpus with split German and unchanged English. Since in this corpus “Grund” is often broken off from a compound, the compound splitting module learns the translation table entry “Grund”→“basic”. By joining the two translation lexicons, the same method may be applied, but this time with the correct split of “Grundrechte”.

A vast amount of splitting knowledge (for this data, 75,055 different words) is acquired by splitting all the words on the German side of the parallel corpus. This knowledge contains for instance that “Grundrechte” was split up 213 times and kept together 17 times. When making splitting decisions for new texts, the compound splitting module 120 may use the most frequent option based on the splitting knowledge. If the word has not been seen before, the compound splitting module may use the frequency method as a back-off.

The POS module 225 may be used to prevent errors involving the splitting off of prefixes and suffixes. For instance, the word “folgend” (English: “following”) may be broken off into “folgen” (English: “consequences”) and den (English: “the”). This occurs because the word “the” is commonly found in English sentences, and therefore taken as evidence for the existence of a translation for “den”. Another example for this is the word “Voraussetzung” (English: “condition”), which is split into “vor” and “aussetzung”. The word “vor” translates to many different prepositions, which frequently occur in English.

To exclude these mistakes, the POS module 225 may only break compounds into content words, e.g., nouns, adverbs, adjectives, and verbs, and not prepositions or determiners. The German corpus may be tagged with POS tags using a tagger, e.g., the TnT tagger, which is described in Brants, T., “TnT—a statistical part-of-speech tagger,” Proceedings of the Sixth Applied Natural Language Processing Conference ANLP (2000).

The POS module 225 may obtain statistics on the POS of words in the corpus and use this information to exclude words based on their POS as possible parts of compounds.

Experiments were performed using a corpus of 650,000 NP/PPs. The corpus included an English translation for each
In the columns, “correct-split” refers to words that should be split and were split correctly. “Correct-not” refers to words that should not be split and were not split. “Wrong-split” refers to words that should be split but were not split. “Wrong-faulty” refers to words that should be split, but were incorrectly (either too much or too little). “Wrong-split” refers towards that should not be split, but were split. “Precision” is the ratio of (correct split)/(correct split+wrong faulty split+wrong superfluous split). “Recall” is the ratio of (correct split)/(correct split+wrong faulty split+wrong not split). “Accuracy” is the ratio of (correct)/(correct+wrong).

In the rows, “raw” refers to the results with unprocessed data with no splits. “Eager” refers to the biggest split, i.e., the compound split into as many parts as possible. If multiple biggest splits are possible, the one with the highest frequency score is taken. In the “frequency based” method, the word is split into most frequent words. In the “parallel” method, the split is guided by splitting knowledge from a parallel corpus. In the combined “parallel and POS” method the split is guided by splitting knowledge from a parallel corpus with an additional restriction on the POS of split parts.

For one-to-one correspondence, the most sophisticated method that employs splitting knowledge from a parallel corpus and information about POS tags provides the best results, with 99.1% accuracy. The main remaining source of error is the lack of training data. For instance, the method failed on more obscure words such as “Passagier-autokommen” (English: “passenger volume”), where even some of the parts have not been seen in the training corpus.

An experiment was performed to test translation quality with a word-based MT system. The translation model used was the IBM Model 4. The system was trained on the 650,000 NP/PPs with the Giza toolkit, and the translation quality was evaluated on the same 1000 NP/PP test set as in experiment described above for one-to-one correspondence. Training and testing data was split consistently in the same way. The translation accuracy is measured against reference translations using the BLEU score, described in Papineni et al., “BLEU: a method for automatic evaluation of machine translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) (2002). The results are shown in Table 4.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>0.291</td>
</tr>
<tr>
<td>Eager</td>
<td>0.222</td>
</tr>
<tr>
<td>Frequency based</td>
<td>0.317</td>
</tr>
<tr>
<td>Parallel</td>
<td>0.294</td>
</tr>
<tr>
<td>Parallel and POS</td>
<td>0.306</td>
</tr>
</tbody>
</table>

In this experiment, the frequency based method produced better translation quality than the more accurate methods that take advantage of knowledge obtained from the parallel corpus. One reason for this may be that the system recovers more easily from words that are split too much than from words that are not split sufficiently. However, this has limitations as shown by the poor results of the eager method.

Compound words violate the bias for one-to-one word correspondences of word based statistical MT systems. This is one of the motivations for phrase based systems that translate groups of words, such as that described in co-pending application Ser. No. 10/402,350, filed Mar. 27, 2003, which is incorporated herein in its entirety. The results are shown in Table 5.

The translation quality was also tested using a phrase-based MT system. This system was trained with the different flavors of the training data, and the performance was evaluated as before.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>0.305</td>
</tr>
<tr>
<td>Eager</td>
<td>0.344</td>
</tr>
<tr>
<td>Frequency based</td>
<td>0.342</td>
</tr>
<tr>
<td>Parallel</td>
<td>0.330</td>
</tr>
<tr>
<td>Parallel and POS</td>
<td>0.326</td>
</tr>
</tbody>
</table>

Here, the eager splitting method that performed poorly with the word-based statistical MT system gave the best results. The task of deciding the granularity of good splits may be deferred to the phrase-based statistical MT system, which uses a statistical method to group phrases and rejoin split words. This turns out to be even slightly better than the frequency based method.

In an embodiment, the words resulting from compound splitting could also be marked as such, and not just treated as regular words.

A number of embodiments have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the invention. For example, blocks in the flowchart may be skipped or performed out of order. Accordingly, other embodiments are within the scope of the following claims.

The invention claimed is:

1. A machine translation system implemented method, the method comprising:
identifying one or more split options for a compounded word in a source language, each split option having a translation in a target language, the identifying performed by the machine translation system;
ranking the compounded word and the one or more split options, the ranking performed by the machine translation system; and
selecting a translation option from the compounded word and the one or more split options, the selecting performed by the machine translation system.

2. The method of claim 1, further comprising: providing the translation option to a machine translation system for translation into the target language.

3. The method of claim 1, wherein said ranking comprises: ranking the compounded word and the one or more split options based on the number of split options.

4. The method of claim 1, wherein said ranking comprises: ranking the compounded word and the one or more split options based on the frequency of occurrence of the compounded word and the one or more split options in a source language corpus.

5. The method of claim 1, wherein said identifying comprises: identifying a translation pair including the compounded word in a parallel corpus, said translation pair including a translation of the compounded word in the target language; and comparing the compounded word and the one or more split options to the translation of the compounded word in the target language.

6. The method of claim 1, wherein said identifying comprises: excluding a potential split option based on a part-of-speech of said potential split option.

7. The method of claim 6, wherein the part-of-speech comprises one of a preposition and a determiner.

8. An apparatus comprising:
   a split generator to identify one or more split options for a compounded word in a source language, each split option having a translation in a target language;
   a module to generate ranking information for the compounded word and the one or more split options; and
   a split selector to rank the compounded word and the one or more split options based on the ranking information and select a translation option from the compounded word and the one or more split options.

9. The apparatus of claim 8, wherein the module comprises:
   a frequency module to identify the frequency of occurrence of the compounded word and the one or more split options in a source language corpus.

10. The apparatus of claim 8, wherein the module comprises:
    a translation lexicon to identify a translation pair including the compounded word in a parallel corpus, said translation pair including a translation of the compounded word in the target language, and compare the compounded word and the one or more split options to the translation of the compounded word in the target language.

11. The apparatus of claim 8, wherein the module comprises a translation table generated by splitting compounded words in a parallel corpus and aligning the split compounded words with corresponding target words in the parallel corpus.

12. The apparatus of claim 8, wherein the module comprises:
    a module to exclude a potential split option based on a part-of-speech of said potential split option.

13. The apparatus of claim 12, wherein the part-of-speech comprises one of a preposition and a determiner.

14. An article comprising a machine-readable medium including machine-executable instructions, the instructions operative to cause a machine to:
    identify one or more split options for a compounded word in a source language, each split option having a translation in a target language;
    rank the compounded word and the one or more split options; and
    select a translation option from the compounded word and the one or more split options.

15. The article of claim 14, further comprising instructions to cause the machine to:
    provide the translation option to a machine translation system for translation into the target language.

16. The article of claim 14, wherein the instructions for ranking comprise instructions to cause the machine to:
    rank the compounded word and the one or more split options based on the number of split options.

17. The article of claim 14, wherein the instructions for ranking comprise instructions to cause the machine to:
    rank the compounded word and the one or more split options based on the frequency of occurrence of the compounded word and the one or more split options in a source language corpus.

18. The article of claim 14, wherein the instructions for ranking comprise instructions to cause the machine to:
    identify a translation pair including the compounded word in a parallel corpus, said translation pair including a translation of the compounded word in the target language; and
    compare the compounded word and the one or more split options to the translation of the compounded word in the target language.

19. The article of claim 14, wherein the instructions for identifying comprise instructions to cause the machine to:
    exclude a potential split option based on a part-of-speech of said potential split option.

20. The article of claim 19, wherein the part-of-speech comprises one of a preposition and a determiner.