SoSe 06 - Project „Machine Translation“ - Part II

Example Based Machine Translation - Pattern Extraction - part II

Cristina Vertan, Walther v. Hahn

Pattern Extraction

- Monolingual phase
  - find the longest collocation sequences, independently in SL and TL (based on a co-occurence in minimal 2 SL/TL sentences)
- Bilingual phase:
  - Global alignment of collocations (which collocation in SL correspond to which one in TL)
  - Construction of patterns
  - Alignment of text fragments

Bilingual Phase - Task

- SL and TL collocations extracted in the monolingual phase (monolingual patterns) are „aligned“ on the basis of co-occurence criteria,
- Alignment in this phase means that for each collocation in SL exactly one correspondence in TL is found,
- For the moment no alignment between parts within the collocations is performed.

Bilingual Phase - Rationale of the algorithm

- The effectiveness of EBMT lies in the retrieval of the longest possible matching sequences (in the input and translation database),
- The longer the matching, the smaller the chance to find more than one translation equivalent for one SL sequence in the DB
- Therefore we consider only leaf nodes in the collocation trees obtained in phase 1,
- The leaf nodes represent the longest possible word sequence in SL /TL,
- SL and TL strings that co-occur in 2 or more sentence pairs are considered to be translations of each other.

Bilingual phase - Algorithm

- Take each leaf node in the SL/TL collocation trees (obtained in phase 1) represented by:
  - The words contained in the collocations,
  - A list of sentence IDs in which each collocation appears
- Align those SL and TL collocations, which share exactly the same ID-list (according to the monolingual phase a leaf node has at least 2 sentences associated),
- Scan the SL and TL sentences associated to the aligned collocations and built the patterns as follows:
  - Words in the collocation are fixed parts in the pattern,
  - The text between two words in a collocation is a variable in the pattern,

Bilingual Phase - Example

- (gave)(up) Sentence IDs [1,2]
- (habe)(aufgegeben) Sentence IDs [1,2]
- (habe) (verlassen) Sentence IDs [1]

- Aligned collocations (gave)(up) ↔ (habe)(verlassen)

Pattern:

- (...) gave (...)up ↔ (...)(habe)(.....)(verlassen).

Bilingual Phase - possible problems

- Constructing patterns with inflected patterns reduce the generality of patterns, i.e.
  - A sentence containing “(give)(up)” will not be matched with the pattern (without further pre-processing steps),
  - What to do with cases like:
    - (gave) (up) Sentence IDs[1,2,3,4] ,
    - (habe)(verlassen) Sentence IDs [1,2],
    - (haben)(verlassen) Sentence IDs [3,4].
- To explore in the project
  - if such situations occur in our project, and if, how frequently,
  - make a list of incorrectly aligned patterns, or situations in which an alignment is not possible,
  - apply the morphological component and work with word stems
  - compare the obtained results, with and without a morphological analyzer.

Alignment of Text Fragments and Variables

Problem:

After the bilingual phase we have correctly aligned only the fixed parts of the patterns,

There is a common understanding, that the variables (called text fragments) are also translations of each other, however,

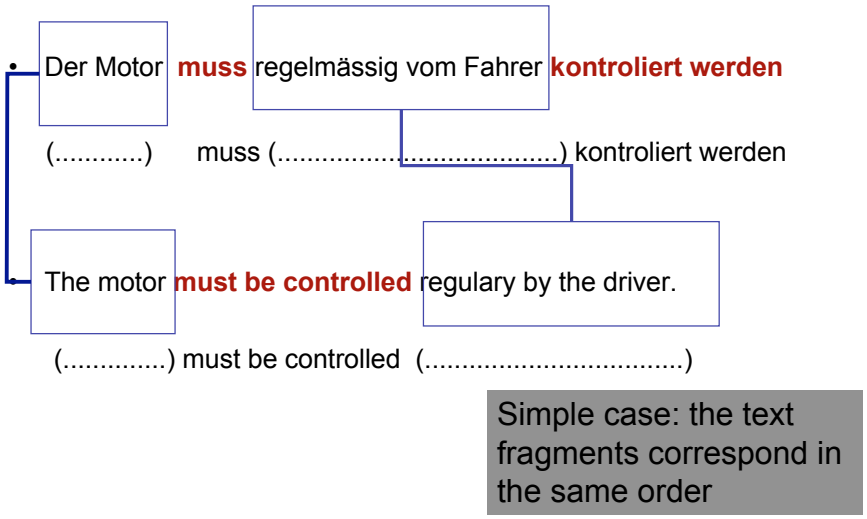
We do not know a priori, which text fragments in the SL sentence is equivalent to which fragment in the TL sentence,

Therefore, the algorithm must find bijective (1:1) and non-bijective relationsships of the type m:n, where $m \neq n$

Alignment of Text Fragments and Variables vs. Sentence alignment

- In sentence alignment it is sufficient, that the SL and TL sentence share the same lexical items,
- when aligning sequences inside sentences we must take into account, that the order of words or subsentential text fragments between two languages are often dissimilar,
- the algorithm must also handle non-adjacent alignments in order to compute long-distance dependencies.

Alignment of Text Fragments and Variables -Example 1-



Simple case: the text fragments correspond in the same order

Alignment of Text Fragments and Variables -Example 2-

- Günstige **Diesel Filter werden in VW-Läden** angeboten
- **VW-shops are selling** unexpensive **Diesel filters**.
- **Diesel Filter werden in VW-Läden** zu günstigen Preisen angeboten

Different translations inferred for the same word

Alignment of Text Fragments and Variables -Example 3-

- Ethiopia was supplied regularly with aid by France
- L'aide était fournit régulièrement à l'Ethiopie par la France

Different order in the alignment of text fragments

Alignment of Text Fragments and Variables - Approach-

1. Compute initial alignments assuming that all local alignments are adjacent (using Dynamic Programming) (similarity measure edit distance),
2. Compute the set of possible non-adjacent alignments (similarity measure : bilingual similarity score),
3. If any non-adjacent alignments are computed, they are recorded and removed from the two sequences, which are then realigned as in step 1,
4. The final global alignment is a concatenation of the non-adjacent alignments and the sequences determined in step 3,
5. If no non-adjacent alignments were computed in step 2, step 3 is not applied, and the final global alignment consists of the alignments determined in step 1.

Alignment of Text Fragments and Variables - bilingual lexical distribution (BLD) -

- works with **cognates**= identical meaning and similar word forms across languages, (i.e for DE-EN : Apfel, Bär, Morgen, hundert, kommen)

Principle:

- given a bilingual corpus aligned at the sentence level
- S = set of SL sentences containing the SL fragment
- T = set of TL sentences containing the TL fragment

$$BLD = 2(|S \cap T|) / (|S| + |T|)$$

Alignment of Text Fragments and Variables - bilingual lexical distribution (BLD) refinement -

- Define manually stop-lists in each language = lists of very frequent words (e.g. conjunctions),
- Remove them from the text fragments,
- Compute BLD on the new text fragments,
- However, this makes the similarity metric language dependent.

Alignment of Text Fragments and Variables - bilingual similarity metric -

- It is a combined score based on the number of cognates shared by the text fragments and the similarity of the distributions of the text fragments (BLD)
- BS = (BLD + |Cognates|) / (1+ |Cognates|)
- In this formula the cognates play a very important role.
- Depending on the language pair, the formula can be modified
- Cognates can be determined with Levenshtein distance

Internet sources

Examples of German-English cognates

- http://www.geocities.com/CollegePark/Classroom/2927/cogs.htm

Overview of similarity measures:

- http://www.dcs.shef.ac.uk/~sam/stringmetrics.html