Experiments with matching algorithms in example based machine translation

Cristina Vertan, Vanessa Espin Martin
University of Hamburg, Natural Language Systems Division
vertan@informatik,uni-hamburg.de

Abstract:

In this paper we will present several matching algorithms used for an example-based machine translation system between English and Spanish. The translation database was extracted from the Web and transformed accordingly fort he purposes of the system. We will describe how a string-based matching algorithm can be improved through the use of morphological and semantic information.

1. Introduction

Example based machine translation is a variant of corpus based MT, and is based on the following ideas.

- Humans do not translate a simple sentence by doing deep linguistic analysis rather,
- Humans translate by properly decomposing an input sentence into certain fragments. The translation of each fragment is then performed by the analogy translation principle, via proper examples.

An example based machine translation (EBMT) system retrieves similar examples (pairs of source phrases, sentences, or texts and their translations) from a database of examples (translation memory).

An EBMT System has the following features (Sumita and Iida, 1990):

- It is easy to be upgraded, by adding appropriate examples to the data-base
- It assigns a reliability factor to the translation result
- It is accelerated effectively by both indexing and parallel computing
- It is robust because of best-matching reasoning
- It uses translator expertise

The main operations to be performed in an EBMT – system are the following:

- Matching. The input is matched against to the database of examples. Examples similar or identical with the input are retained.
  - Alignment: fort he retrieved chunks the corresponding translations are retrieved
  - Recombination: the translated chunks are rearranged to construct a correct sentence in the target language (Way and carl, 2003)

In this paper we will present experiment, which we performed on matching algorithms for an English-Spanish example, based machine translation system. We extracted the corpus from the web, and adapted in order to serve for our purposes. We present two matching methods, one string-based and one on semantic relations.

2. Preparing the corpus (translation memory)

Our corpus was extracted from the Web site http://www.spain.info and contains texts about tourism in Spain in two different languages Spanish and English. The corpus contains approx. 10 0000 words. It contains information about Spanish regions, Spanish society, Spanish touristy routes Following pre-processing operations were performed on the corpus:

a) Document /alignment: texts were separated in Spanish and English documents
b) Sentence alignment: each sentence in one language was linked with is correspondent in the other language
c) Tagging: done in XML the tags relate not only sentences but also noun, verbs and adjectives in both languages.

Initially the corpus contains 241 sentences in Spanish and 223 sentences in English

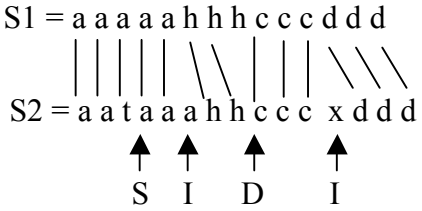
It is important to be mentioned the alignment was done manual, as the texts were not identical translation one of the other

3. Matching algorithms

3.1. Edit distance

The edit distance between two strings `s1` and `s2` is defined as the minimum number of operations performed in order to obtain `s2` from `s1` (Levenshtein algorithm). The measure gives an indication for the closeness of the two strings. Usually the operations considered are: insertion, deletion and substitution

In figure 1 we illustrate how the edit distance between two strings is computed.



Edit_dist(S1,S2) = 4

Figure 1. Edit Distance

In case of EBMT we use the edit distance in order to compare the input sentence with different examples in the translation memory. Two problems appear when using this method:

a) It measures differences between strings and not words, which means that words which are semantically very close (for example inflected forms) could be measured by the edit distance quite different. The problem is more complicated in case of synonyms, which are judged as complete different by the edit distance. Solution to this problem will be presented in section 2.3

b) When the translation memory is built from a parallel corpus, the constituents are quite big sentences. Let us take the following example:

Translation memory: „*Design is he key player in the Spanish fashion industry, which in recent years has gone from strength to strength abroad, thanks to the work of creators such as Jesus del Pozo, Adolfo Dominguez, Paco Rabanne, Pedro del Herro, and the increasing presence of Spanish models on international catwalks"*

Input: *„The key player in the Spanish fashion industry is mode design"*

If we apply the edit distance we will obtain a result greater than 100, although the input sentence is almost complete contained in the translation memory example. The problem is hat only the sentences in the translation memory having an edit distance under a specified threshold are retained for further processing Steps. This threshold cannot be set up so high in order to cover cases as the above mentioned one.

Therefore we transformed the examples in the translation memory as follows:

- Colons, semicolons and commas were used as chunks separators
- A sentence may contain only one verb (as far as possible)
- 1 to 1 sentence alignment (of course there are cases with 1 to 0 or 0 to 1 alignment)

With these transformations our corpus contains, in the second version: 507 Spanish sentences and 503 English sentences, each sentence containing approx. 9 words. We fixed the threshold fort he edit distance at 20, estimated as approx. the double of the number of words/sentence.

3.2. Word by word Matching with semantic information

In contrast with the previous 2 methods, this procedure is not based on comparison of sequence of characters but of words. The procedure looks fort he most similar chunks in the translation database. Similarity is measured according to a thesaurus containing concepts in the translation memory and the associated words.

Following resources were used:

- Translation memory
- Bilingual dictionary
- Thesaurus.

a) The Thesaurus

A thesaurus is a set of terms building a vocabulary of semantically related terms, which covers a specific domain of knowledge. For our experiments we built a bilingual thesaurus of nouns. The structure of this thesaurus is shown in figure 2:

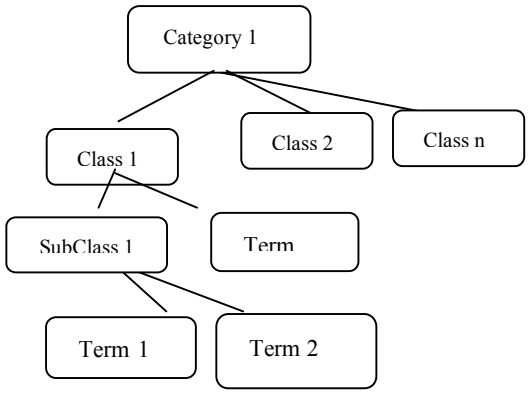


Figure 2. Thesaurus architecture

In figure 2 we show the structure of one of the possible categories: each category can be further divided in subclasses. The leafs are corpus nouns (terms). Terms can belong to subclasses, classes or categories. Following excerpt from our thesaurus corresponds to the structure in figure 2:

Spanish culture
    Entertainment
        Fashion
        Sports
    Religion
    Dietary Habits
        Mediterranean Diet
            Typical Food
            Tapas
    Art
        Monuments
            Mosque
            Churches
            Museum
            ....
    ....

Proper nouns are not supposed to appear in the thesaurus, but it is important to provide a way to relate one each other. We ensured this by adding a specific Attribute in the tag corresponding to the proper noun. This attribute specifies a related concept in the thesaurus.

Example:

<PN type="map">Andalusia</PN> (corresponding to „territories" class in the thesaurus).

<PN type= „monum">Alhambra>/PN> (Corresponding to „monuments" class in the thesaurus)

Terms were related through following WordNet – like relations:

- Preferred terms / non-preferred terms: the preferred terms are those terms used for indexing; the non-preferred terms belong to the corpus and can be referred and are synonyms of a preferred term
- Synonyms: the synonyms of a term
- UF. Used for. The terms that have these relations are those preferred terms used for indexing another terms, for example synonyms
-USE: the terms having this relation are non-preferred terms. The tag indicates which terms are suitable for indexing when this term is needed

- Translation: Translation of the term
- Language

b) The bilingual dictionary

We developed a full-form lexicon. For each word we specify the stem and additional morphological information 8POS, gender, number, etc.).

In a first Step we annotate the corpus in the following way:

- For terms not belonging to the thesaurus:

    <verb lex = „border"> borders >/verb>

- For terms connected with concepts in the thesaurus:

    <N lex="river">rivers>/N>

The lexicon is built at the compilation time, i.e. terms are automatically extracted from the corpus and alphabetically ordered, e.g.:

```
<term>
      <word>coruña </word>
      <lema>geo</lema>
</term>
<term>
      <word>ancatilados</word>
      <lema>ancatilado>/lema>
</term>
</term>
.....
```

In order to compute the distance following are performed:

- Lemma extraction from the corpus
- Transform the translation memory by substituting each word by its lemma

On the transformed translation memory the distance is computed according to the formula

(1) $dist = (I+D+ 2*\Sigma semdist) / (Length\_input + Length\_example)$

The number of insertions (I), deletions(d) and substitution (S) operations are summed up and the total is normalized by the sum of the length of the example and input sequences.

Substitution is calculated as the semantic distance between two substituted words. Semdist is defined as the division of K (the level of the least common abstraction in the thesaurus of the two considered words) by N (the height of the thesaurus) (Sumita and Iida, 1991)

4. Conclusions and further work.

In the previous paragraphs we shown how string-based matching can be improved by adding syntactical and semantic information. In this way we are able to retain for further processing also examples which are not identical with the input. For the moment we considered only nouns fort he thesaurus construction, and the morphological

annotation is done only for nouns and verbs. We intend to extent the morphological annotation to all POS, and to develop broader the thesaurus. We preview also the extension of the thesaurus with more relationships.

References:

Charles Meyer, 2002, 2english corpus Linguistics. An introduction“, Cambridge University Press

Andy Way and Michael Car, 2003: „Recent Advances in Example Based Machine Translation“, Kluwer Academic Press

Eichiro Sumita and Iida, 1991, „Experiments and Prospects of Example-Based Machine Translation“, Proceedings of ACL-91, pp. 185-192

Eichiro Sumita, 2001, „Example-base machine translation using DP-matching between sequences“, DDMT workshop of 39th ACL, pp 1-8