Lab 4: Machine Translation

TDT4275: Natural Language Interfaces

Jonas Myrlund

April 16, 2013

1 Written Assignments

I chose the two online machine translation products PROMT[1] and Google Translate[2]. I chose to use the French sentence "Mon aéroglisseur est plein des anguilles, et donc j'ai besoin de pratiquer mon français." for this exercise.

The results from the two translators are shown below:

PROMT	*My hovercraft is full of eels, and therefore I need to play my French.*
Google Translate	*My hovercraft is full of eels, so I need to practice my French.*

Figure 1: Comparison of machine translations.

Based on this, I would go for the following translation: "My hovercraft is full of eels, and so I need to practice my French." Most notably, I've gone for a combination of the translations "and therefore" and "so" after the comma.

The two systems have chosen slightly different ways of translating *et donc*: Google Translate has chosen a simple "so", while PROMT is more verbose with its "and therefore". However, they both work nicely in the context of the sentence, so we'll let both pass.

The translation of *pratiquer*, however, is obviously completely off in the PROMT translation, the wrong sense of the word, "play" having been chosen. Google Translate has chosen "practice", which seems like a much better thing to do with "French".

Due to the error in the word sense of *pratiquer*, the Google Translate-version seems the better translator in this case.

[1] `http://translation2.paralink.com/`

[2] `http://translate.google.com/`

2 Building an SMT system

I'll be using two different corpora for my two different sets of languages.

Spanish-English	EU parliament corpus
Finnish-Norwegian	Movie subtitles corpus

A

Explain why is it important to set a side a separate corpus for evaluation.

When training an agent, there is always a degree of overfitting. That is, the agent will be better suited at solving tasks found in the training data than tasks it has never seen before. The only way to lower the degree of overfitting is to use a large enough training corpus.

However, overfitting *will* occur; a test set is our way of knowing to what extent.

Is there a way to use the entire corpus in training and evaluation while still adhering to these concerns?

No. What makes a test set useful in the first place is that it is *independent* of the training data, while it follows the same probability distribution.

This is also the main reasoning behind why we split each utilized corpus into a training and a test set, instead of having a single test set used for testing various other corpora.

B Use your prepared corpus to build a model.

Build two such models for your chosen language pairs. Show the output of the model directory and report the number of entries (lines) in the phrase table.

The model directories contain the following files:

- aligned.grow-diag-final
- extract.sorted.gz
- lex.f2e
- phrase-table.gz
- extract.inv.sorted.gz
- lex.e2f
- moses.ini

The phrase tables have the following number of entries:

- Finnish-Norwegian phrase table: 264,500 entries.
- Spanish-English phrase table: 67,268 entries.

As is evident, my corpus sizes are quite small, as it turned out running the `train-model.perl` took ages with my original corpus sizes.

C Decoding.

Use your newly create models to create translations for your test sets. You might notice that decoding multiple sentences is almost as fast as training one. Why is that? The computational complexity is much larger for training than it is for decoding. Furthermore, when decoding, the models need only be read from disk, not modified.

Try a couple of out-of-domain sentences, for example, about food or sports. How does your system treat these sentences? In the case of Finnish to Norwegian, this turns out to work very poorly. As both Finnish and Norwegian use compound words to a large extent, many words were not recognized, resulting in very poor translations.

An example of this is shown below (*emphasis* indicates unknown words):

Finnish	Ilmatyynynalukseni on täynnä ankeriaita
Norwegian	*Ilmatyynynalukseni* er full av *ankeriaita*

My other two languages, English and Spanish, aren't as prone to using compound words as Norwegian and Finnish, but the choice of corpus still ensured that a large number of common enough words went unrecognized. The corpus, as mentioned, is from the European parliament documents, and fared extraordinarily poorly when consulted with the sentence "a vanilla ice cream is packed full of calories". It recognized neither *vanilla*, *ice* or *cream*, and ended up with the sentence: "un vanilla ice cream es packed pleno de calories".

D Evaluation.

Evaluate the output from you system with the script multi-bleu.perl bundled with Moses. Be sure to evaluate on the test corpus. How do you interpret the results? Skim through the target language output of both the closely and the distantly related languages. Which do you think is best yourself? Report the BLEU scores. The evaluation of the *distantly related* languages Norwegian and Finnish yield the following BLEU result:

```
BLEU = 67.93, 80.7/74.7/70.9/66.4
```

```
  (BP=0.931, ratio=0.933, hyp_len=25125, ref_len=26918)
```

The evaluation of the more closely related languages English and Spanish yield the following:

```
BLEU = 56.60, 71.8/58.6/52.2/46.9
  (BP=0.999, ratio=0.999, hyp_len=3779, ref_len=3783)
```

Both language pairs perform quite poorly, with a score of only 67.9 and 56.6, respectively. A larger or more generic corpus would probably benefit both cases vastly.

Although the languages are more distantly related, Norwegian and Finnish outscore Spanish and English. This is probably due to the choice of corpus; movie subtitles tend to have quite shorter and less complicated sentences than parliamentary documents.