Machine Translation and Parallel Corpora

University of Zurich, May/June 2012
Martin Volk, Magdalena Plamada

Final Task: Lessons and Visions

This final task serves two purposes. First, we would like to check what you understood of the material presented and discussed in class and second we would like to hear / read about your vision of the field in the future.

1. In this course we have built **Statistical Machine Translation systems for subtitles** based on freely available subtitles.
    a. Why are these systems not as good as the ones reported in [Volk 2010]? Give three reasons.
    b. How could you improve the translation quality of your system by filtering the training corpus?
    c. How could you reduce the number of unknown words for your SMT system?
    d. We have argued that TV subtitles are particularly well suited for Machine Translation. What were the arguments? What other **text types** are similarly well suited?

2. As task 2 we have programmed **IBM model 1.** Describe briefly how it works and how it differs from IBM model 2. What are the highest translation probabilities after 5 iterations of the IBM model 1 algorithm for the words *Haus* and *sein* when we have only the following two sentence pairs as input (assuming that we start with a uniform probability of 0.5):

    sein Haus ist klein --- his house is small
    das Haus kann klein sein --- the house can be small

    Explain the observed translation probabilities.

    What happens if we add the following as the third sentence pair?

    sein Auto ist klein --- his car is small

    How do the translation probabilities for *Haus* and *sein* change?

    What happens if we apply a Part-of-Speech tagger on the German side before word alignment and then use the following three sentence pairs instead of the above as input to the IBM model 1 algorithm?

    sein_PPOSAT Haus_NN ist_VAFIN klein_ADJD --- his house is small
    das_ART Haus_NN kann_VMFIN klein_ADJD sein_VAINF --- the house can be small
    sein_PPOSAT Auto_NN ist_VAFIN klein_ADJD --- his car is small

    [The PoS tags are from the STTS, Stuttgart-Tübingen Tag Set.]

3. What is the role of the **language model** in Statistical Machine Translation? Why is it sometimes useful to train a language model over Part-of-Speech tags rather than words? Why do we use smoothing when training a language model?

4. We discussed a method for combining statistical and linguistic information which is called **Factored Machine Translation**. Briefly sketch how this works.

   Another method for combining statistical and linguistic information is called **hybrid machine translation** which combines rule-based MT (= linguistically driven MT) and SMT. One such method for English-Spanish MT is described in the paper

   > Victor M. Sanchez-Cartagena, Felipe Sanchez-Martinez, Juan Antonio Pérez-Ortiz: *Integrating shallow-transfer rules into phrase-based statistical machine translation*. In Proceedings of the 13th Machine Translation Summit, p 562-569, September 19-23, 2011, Xiamen, China.
   > http://www.dlsi.ua.es/~fsanchez/pub/pdf/sanchez-cartagena11c.pdf

   Summarize the method in this paper and discuss the results.

5. Let us assume that ABB hires you for consulting about Machine Translation. They have user manuals for their photovoltaic systems. They need to translate these user manuals from English to German. For this purpose they have purchased a rule-based machine translation system. They notice that this system makes mistakes and wonder if it is better to use this rule-based MT system or Google Translate. How would you set up a **systematic comparative evaluation** of the two MT systems for ABB? How much will your service to plan and to perform the evaluation cost?

6. What is your **vision on Machine Translation in mobile devices**? In which application scenarios (which services, which languages, which mode) will Machine Translation be particularly useful?
   Here we are interested to learn what you think will happen (a realistic scenario) and also what you think will probably not happen but would be nice if it happened (a wish list scenario). This is your chance to share your fantasy. ☺ Please give us enough arguments to believe in your predictions.

Important: Each student shall answer this final task individually. Help from other people must be acknowledged. Quotes from books or from electronic sources must be marked as such and must be accompanied by a complete reference to the source.

To be delivered:

Please write at least one paragraph on each of the above questions so that your overall report adds up to **about 4 pages**. Please name your report **Your_Name_Final_Task.pdf**

Deadline: Please submit your report to OLAT before **Monday, 25. June 2012, 12.00h**.