Machine Translation and Parallel Corpora

University of Zurich, May/June 2015
Martin Volk, Mathias Müller

Final Task: Lessons and Visions

1. In Statistical Machine Translation we typically perform automatic **word alignment** before **phrase alignment**.

    a. [1 points] Why does our algorithm work in this order?
    b. [1 points] Why do we use symmetrization on word alignment before computing phrase alignment?
    c. [8 points] Perform the following experiment: Use 1 million words of the German-French parallel portion of the Text+Berg corpus (Release 151 v01; find access information below) according to your own selection criteria. Please describe them.

    Train the word alignment with GIZA++ as you did in exercise 3 in our course. Investigate the resulting symmetrized word alignment according to *grow-diag-final-and* for 3 sentences that contain different translation variants of the German word *Höhe*. Present your word alignments as alignment matrices. You may use the provided tools in OLAT for automatically creating alignment matrices if you like. How many correct and incorrect alignment points do you observe in your 3 matrices?

2. In this course you have built **Statistical Machine Translation systems for subtitles** based on freely available subtitles.
    a. [2 points] How could you improve the translation quality of your system by automatically filtering the training corpus (ie. ignoring part of the training corpus)?
    b. [2 points] How could you reduce the number of unknown words for your SMT system?
    c. [2 points] How could you get your SMT system to pick a better translation from the n-best list than the top one chosen by the decoder?

3. In this course we have learned that the correct translation of **connectives** (e.g. conjunctions like "although, since, while") is important for the comprehensibility of machine translation output. Unfortunately, the experiments on improved translation of connectives presented by Thomas Meyer did only lead to small improvements in BLEU scores.
    a. [3 points] Describe briefly three experiments that Thomas Meyer performed.

b. [2 points] Why did the improvements have only a little impact on the BLEU scores? Which other means of automatic MT evaluation would be necessary to capture the improved translation of connectives?

4. One important research area in SMT is "**Automatic Quality Estimation**" with the purpose to suppress bad translations.

    a. [4 points] How does Automatic Quality Estimation work as described by

    Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, Nello Cristianini (2009): Estimating the sentence-level quality of machine translation systems. In: *Proceedings of 13th Conference of the European Association for Machine Translation*. p. 28-37

    b. [1 point] Why is Automatic Quality Estimation for SMT much more difficult than Computing Fuzzy Match scores in Translation Memory systems?

    c. [1 point] Why are professional translators interested in having quality estimation scores on SMT output?

5. The (fictional) pharmaceutical company Vonartis wants a **Trilingwis** system as online dictionary with usage examples over their document collection English – French – German. They offer texts that sum up to 10 million words in German with translations in English. But only 6 million German words are translated into French.

    a. [4 points] Please specify the steps necessary to build the Vonartis Trilingwis system.
    b. [2 points] How will you deal with the differences in corpus size between English, French and German when computing the frequencies of the translation variants and displaying them together?

6. Let us assume that Credit Suisse hires you for consulting about Machine Translation. They have internal communication documents, customer information brochures and various versions of "Terms and Conditions" which they need translated from German to French and Spanish. For this purpose they have purchased one rule-based machine translation system for each language pair. They notice that these systems make mistakes and wonder if it is better to use Google Translate or custom-made SMT systems rather than these rule-based MT systems.

    a. [5 points] How would you set up a **systematic comparative evaluation** of the rule-based and statistical MT systems for Credit Suisse?
    b. [2 points] What is a likely set of recommendations that will come out of this evaluation?

Important: Each student shall work on this final task individually. Help from other people must be acknowledged. Quotes from books or from other printed or electronic sources must be marked as such and must be accompanied by a complete reference to the source. In case we have any doubt whether the submitted solutions originate from a student, we reserve the right to invite the student to an oral exam.

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, 15. June 2015, 18.00h**.

How to obtain the Text+Berg corpus?

You will find our latest Release_151_v01 of the Text+Berg corpus (151 years; from 1864 to 2014; > 500 MByte) at

http://kitt.cl.uzh.ch/kitt/textberg/release/Text+Berg_Release_151_v01.zip

user: textberg
password: tbaccess

The README file that comes with the release explains the corpus structure and annotation format.