Machine Translation and Parallel Corpora

University of Zurich, May/June 2014
Martin Volk, Anne Göhring

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 you have built **Statistical Machine Translation systems for subtitles** based on freely available subtitles.
    a. [3 points] Why are these systems not as good as the ones reported in [Volk et al. 2010]? Give three reasons.
    Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström (2010). *Machine Translation of TV Subtitles for Large Scale Production.* In: Proceedings of the Second Joint EM+/CNGL Workshop on Bringing MT to the User: Research on Integrating MT in the Translation Industry. Denver. 2010, pages 53-62.

    b. [3 points] How could you improve the translation quality of your system by automatically filtering the training corpus (ie. ignoring part of the training corpus)?
    c. [2 points] How could you reduce the number of unknown words for your SMT system?

2. [2 points] What are Lowercasing and Truecasing in SMT? What are the advantages and disadvantages?

3. As task 2 we have programmed **IBM model 1.**

    a. [2 points] Describe briefly how it works and how it differs from IBM models 2 and 4.

    b. [8 points] What are the highest translation probabilities after 5 iterations of the IBM model 1 algorithm for the words *Auto* and *sein* when we have only the following two sentence pairs as input (assuming that we start with a uniform probability of 0.5). You may use our Python example implementation of IBM Model 1 (in OLAT) if you like. If you use your own implementation, please submit it again with this task.

    er kann sein neues Auto fahren -- he can drive his new car
    das Auto kann neu sein -- the car can be new

-

Explain the observed translation probabilities.

What happens if we add the following as the third sentence pair?
sie kann sein Auto fahren --- she can drive his car

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

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

er können sein neu Auto fahren -- he can drive his new car
das Auto können neu sein -- the car can be new
sie können sein Auto fahren --- she can drive his car

4. One method of adapting SMT systems to particular text types is called domain-adaptation.

    a. [2 points] Why is domain-adaptation important when building SMT systems?

    b. [4 points] How does the domain-adaptation that was developed by Rico Sennrich work?

    Sennrich, Rico (2012). *Perplexity minimization for translation model domain adaptation in statistical machine translation.* In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 2012 - 2012, 539-549. http://www.zora.uzh.ch/61712/

    c. [2 points] How does Sennrich's method differ from domain-adaptation as suggested by

    Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

5. 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 Italian. 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 MT systems for Credit Suisse?
b. [3 points] What is a likely set of recommendations that will come out of this evaluation?

6. [4 points] What is your **vision on the impact of Machine Translation on second language learning**? Will MT systems discourage people from learning a foreign language in the near or medium future? Why should anyone learn a foreign language when MT systems are available everywhere (Lindsay Bywood: "machine translation as a commodity") and translate with a comprehensible quality?

   Here we are interested to read 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 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 **Thursday, 19. June 2014, 18.00h**.

-