Machine Translation Assignment

Your task is to do a mini-research project on some aspect of translation, and write up your findings. Your output should look a bit like an ACL conference paper, i.e., about 8-12 pages comprising

- description/motivation: what is the problem and why it’s important
- a brief literature survey: who’s tackled the same or similar problem, and their findings
- your technique/s for solving the problem
- experimental validation
- concluding discussion

You should seek to do something novel, although this doesn’t mean you have to magic up a brilliant new algorithm in the short time available. You might want to do thorough experimental validation of an existing method, e.g., applying it to new language.

You should think of a topic that interests you, and talk to me about it before beginning work. I’ll help to ensure that it’s a sound idea, is feasible in the time constraints, and give you some pointers to relevant literature.

If you need inspiration, here are some ideas:

- **Translating Twitter.** Standard translation systems struggle on tweets: they’re short so there’s little context, include lots of slang, typos, emoticons etc. This makes translating them into another language tricky. See if you can improve translation performance, e.g., by discovering translations, building better language models or normalizing the text. This is uncharted waters, so you’ll also need a way of evaluating your performance!

- **Transliteration.** Proper nouns are often transliterated into a foreign language rather than translated directly. This usually is a mapping between the phonetic form of the source word to a close match in the target language, although there’s often a fair bit of creativity in the process. You might want to develop a transliteration system from examples, or try doing it without just an orthographic language model, but no parallel training data! **Maths warning: could be quite a tricky undertaking.**

- **Incorporating discourse context into translation.** Translation engines independently translate each sentence, missing many potential clues which could aid translation in the rest of the source document. For instance, it’s rather difficult to translate pronouns (e.g. he/she/it) into a foreign language if you don’t know the intended referent. This project will look at ways to harness information in the surrounding sentences and wider document has the potential to resolve translation ambiguities.

- **Dealing with morphology.** Many languages have complex word forms which incorporate considerable additional information beyond their core meaning. Translating complex words is clearly difficult, especially as we’re unlikely to have seen the word before and therefore will have to split the word into smaller units to even have a chance. Generating morphologically complex forms in the translation output is even more of a challenge. Choose one setting, and see if you can improve the performance!

- **Pivoting.** Translating between language A and B by using an intermediate language has been shown to work rather well. This is despite using a very simple learning strategy -- just using simple frequency estimates. It’s likely that developing a feature-based discriminative classifier could do much better, training the model using a small seed set of known good translations. Features could detect consistent orthographic transforms, syntactic transforms, intermediate words to avoid, and more. **Maths warning: could be quite a tricky undertaking - but there’s a good paper in this!**