1. Paraphrastic Sentence Compression with a Character-based Metric: Tightening without Deletion

In this paper, Napoles et al. present a character-based substitution approach to sentence compression using bilingual parallel corpora. The motivation for this method involves the strict constraints of deletion-based approaches, which leads to the methodological modification of sentences and, overall, abstract and robotic results. Instead, they highlight the use of paraphrases extracted from bilingual corpora and re-ranked using a novel monolingual heuristic to provide more intuitive and naturalistic summarizations of source text, for use in cases such as document simplification, mobile text, subtitling, and micro-blog constraint fitting. At a high level, the intuition behind their methods is such that shorter words can be chosen where possible, freeing up space for more content-filled words. Their results show that combined substitution and deletion compressions preserve more meaning in the same number of characters as deletion-only compressions.

In generating paraphrases, the authors extract paraphrases from bilingual parallel corpora, treating any English phrases that share a common foreign phrase as potential paraphrases of each other. To rank candidates, they propose a two-step process: candidates are first ranked using the translation model probabilities p(e|f) and p(f|e), requiring the paraphrase to be of the same syntactic type, then the candidates are re-ranked using a monolingual distributional similarity metric. This metrics relies on the approximate cosine similarity scores over feature counts of the phrases and has a profound impact on the quality of the paraphrase candidate rankings. Using manual ranking of 1,000 randomly selected paraphrase sets, the addition of the monolingual filtering technique after the original translation score filter introduced a stronger positive correlation between the human rankings and the produced rankings.

To tighten sentences and generate a final compressed form, Napoles et al. use a dynamic programming strategy to find the combination of non-overlapping paraphrases that minimizes a sentence's character length; the monolingual score is not used in any type of weighted score, but can be used as a threshold to ensure a certain level of confidence in the preservation of meaning. To choose between compressions of equal or near-equal length, two metrics are used: the word-overlap score between the original and candidate sentence, and the language model score of the compressed sentence.

In initial evaluations using substitution-only and deletion-only methods, the substitution method performs poorly under all cosine-similarity thresholds between 0.65 and 0.95 at a 0.10 increment. They posit that this is due to erroneous paraphrase substitution of phrases with the same syntactic category and distributional similarity, but different semantics.

In light of this sub-par showing, Napoles et al. then tested the viability of their approach using manual labeling of good paraphrases produced by their model. Comparing substitution-only, substitution-and-deletion, and deletion-only compression methods within 5 characters of difference between their compressions, the methods involving substitution performed better than the deletion-only method in grammaticality and meaning statistics, indicating the potential for substitution-only methods given more advanced paraphrase acquisition and ranking methods.

2. Deep, Multilingual Word Alignments using Cross-Domain Corpora

As noted by the original paper, there are detrimental issues in their chosen method of candidate compression ranking. Although there may be some experimentation on the use of semantic and word-sense disambiguation techniques for filtering out unlikely paraphrases, my project will mainly investigate the combination of multiple cross-domain parallel corpora in improving paraphrastic sentence compressions.

Previous work has focused on using singular bilingual corpora, namely English-German corpora, to test different sentence compression techniques and ranking metrics. However, little work has investigated the use of corpora in different languages and in multiple domains. By using deep-linking word alignments across a variety of corpora and domains, more paraphrases are likely to be found, as single-domain corpora tend towards using the same type of tone and wording. This is especially true of prominent bilingual corpora, as these corpora are often produced as a result of formal international relations and proceedings, and therefore contain more pointed and less varied language.

By combining two or more parallel corpora using linked pairs of word alignments, cross-domain paraphrases can be extracted. This diversifies the wording of the paraphrases, which can be highly beneficial and still fitting depending on the application of compression. For example, news article summarization can benefit from less strict and formal language for posting onto social blogs and websites. In addition, combined corpora can improve the translation and paraphrasing of low-density languages with sparse amounts of parallel data. Phrases in sparse parallel data sets can be mapped into larger domains, revealing a wealth of fitting paraphrases. In addition, there is potential in attempting paraphrastic sentence *decompression*, allowing condensed sentences found on microblogging websites like Twitter and Weibo to be expanded into more standard phrasings. This has direct benefits to linguistic analyses of social media, as the vast assortment of tools traditionally used for linguistic analysis would not need to be modified to fit such a complex domain.

In experimenting with the idea of deep, cross-domain corpora, different techniques will be investigated. At a high level, corpora can be combined in two different ways to extract paraphrases: in parallel, using multiple different English-to-language corpora as per previous work done in the field, and in series, using deep linked alignments to combine domains. The latter will be tested using various depth limits, languages, and language orderings. I theorize that language origins and their orderings within the system will play a large role in the success of deep, multilingual word alignments due to their varied vocabulary sizes and expressions -- some languages will have more specificity within their vocabulary, and a language may express two semantic ideas as one combined expression. Deep word alignments will therefore be reliant on the compression and decompression of their semantic expressions as they are converted from one language to the next.

Many of the experimental specifics will rely on the methods presented by Napoles et al, which in turn relies on the word alignment and phrase extraction methods presented in Koehn et al (2003). A newer word alignment tool like Giza++ or the Berkeley Aligner will likely be used, but I expect to use a rudimentary implementation of Koehn's phrase extraction already available on the web. Once phrase extraction is performed on all chosen pairs of corpora, these phrase mappings can be used to link these alignments. The same ranking metrics will be used, with

some possible additional experimentation in using WordNet semantic information to improve these rankings, if enough time is available after the main experiments.

In terms of the specific corpora that will be used, combinations of parliamentary parallel corpora (e.g. Europarl) and more literary or informal corpora are likely candidates. Recent parallel corpora created for microblogging sites such as Twitter and Weibo have emerged in the past two years, but due to their strict publishing policies, each sentence must be mined locally. Due to this constraint, about 50% of the available corpora is now unavailable due to deleted posts, and it is extremely difficult to mine existing posts due to API request limits. Therefore, experiments involving microblogging corpora will not be possible in the time available.

As already noted, evaluation will include experimentation into corpora combination types (parallel vs. series), different language combinations, language order, depth limits, source/target corpora sizes, and ranking metrics. It will also include baseline metrics using the methods presented in Napoles et al. The original paper uses manual crowd-sourced judging to rank grammar and meaning scores; since I do not have enough time, I will perform manual judging with about 200 sentences. I will also experiment with using probabilistic tools (e.g. the Stanford parser) already used or mentioned in lecture to get rough wide-scale scores.

The goal of this paper will not necessarily be to improve on the best-performing sentence compression methods, but to improve on the best-performing paraphrastic sentence compressions, which has been shown in Napoles et al to have the potential to perform as well or better than current best-performing deletion-only compressions.