Paraphrastic Sentence Compression using Deep-Linking Bilingual Phrase Alignments and Cross-Domain Corpora

Kevin Yeh (kky226)
CS 388L: Natural Language Processing
Spring 2015 | Austin, TX
kevinyeah@utexas.edu

***Abstract*—**This paper presents a new approach to paraphrase acquisition for the purpose of substitution-based sentence compression, which tightens a sentence by reducing its character length. By combining existing paraphrase detection and ranking techniques with cross-domain corpora and deeply-linked phrase alignments across multiple parallel corpora, shorter paraphrases can be discovered to replace source phrases while preserving both meaning and syntactic validity. New paraphrase acquisition methods are introduced that utilize alignments from multiple parallel corpora in a variety of ways. Specifically, the following methodologies are investigated:

1) Combining multiple bilingual parallel corpora in parallel using a ranked distributional metric.

2) Combining multiple parallel corpora in series, using deeply-linked alignments to discover new paraphrases within different domains and languages.

3) Modifying the order of linked alignments based on the class and origin of different languages, implicitly optimizing the compression and decompression of semantic expressions as they are converted from one language to the next.

I Introduction

Sentence compression is the process of shortening a sentence while preserving the most important information in a meaningful and syntactically valid way. As information becomes more accessible through online social mediums, it becomes more necessary for media and social users to be able to convey that information quickly and succinctly while retaining as much of the vital information as possible. In particular, microblogging websites like Twitter, Facebook, and Sina Weibo have become immensely popular forms of public and personal news communication, and these forms of communication have come with strict verbosity limits and natural practicality boundaries.

In addition to social trends in online media, physical and technological constraints require the summarization of text for mobile devices and subtitling. Memory footprints are always of concern for mobile data transfer, and subtitles need to convey accurate information while requiring as little visual attention as possible.

While early sentence compression techniques were developed in support of extractive summarization (Knight and Marcu, 2000)[1], skewing early development towards deletion-based models which extract a subset of words from a longer sentence, more recent work has proven that substitution and character-based techniques have the potential to perform as well, if not better than, deletion-based models. In particular, results from recent paraphrastic sentence compression research (Napoles et al, 2011)[2] show the potential for substitution-only methods given more advanced paraphrase acquisition and ranking methods.

In this paper, ranking methods are not explored, but new methods of paraphrase acquisition are investigated. 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 limiting paraphrase acquisition to a single parallel corpus, the paraphrases tend towards using the same type of tone and phrasal 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.

To remedy this, different techniques involving the combination of multiple cross-domain parallel corpora have been theorized and experimented with, and the results are shown below.

II Previous Work

Barnard and Callison-Burch (2005)[3] presented an initial approach to paraphrasing using bilingual parallel corpora. Extending off of the phrase-based statistical machine translation introduced by Koehn et al. (2003)[4], their method involves aligning phrases within the corpus and equating different English phrases aligned with the same phrase in another language. This assumption of phrase similarity for phrases mapped onto the same foreign phrase is the converse of that made in the word sense disambiguation work of Diab and Resnik (2002)[5].

Since many phrases can be mapped onto the same foreign construct, it is important to also rank the extracted paraphrases using a probability assignment. Bannard and Callison-Burch use a paraphrase probability p(e2|e1) defined in terms of the dual translation probabilities p(f|e1), the probability that an English phrase e1 translates to foreign phrase f, and p(e2|f), the probability that a candidate phrase e2 translates as the foreign language phrase. This probability can be easily calculated using maximum likelihood estimation, formed by counting how often the phrases e and f were aligned in the parallel corpus:

$$p(e|f) = \frac{count(e, f)}{\sum_{e'} e', f}$$

In addition, more contextual information can be used to extend the paraphrase probability by taking into account the sentence, S, that the phrase e1 appears in. Thus, the most probable paraphrase e2' becomes:

$$e_2' = argmax(p(e_2|e_1), S)$$

The work of Napoles et al. (2011) expands on the use of paraphrase extraction to motivate advances in character-based sentence compression in the digital age. The motivation for substitution-based approaches involves the strict constraints of deletion-based approaches, which leads to the methodological modification of sentences and, overall, abstract and robotic results. To counter this, they introduce a new character-based substitution approach using bilingual parallel corpora, highlighting the use of phrases re-ranked using a novel monolingual heuristic to produce more intuitive and naturalistic summarizations of source text, for use in cases such as document simplification, mobile text, subtitling, and micro-blog constraint fitting.

Original	Congressional leaders reached a last-gasp agreement Friday to avert a shutdown of the federal government, after days of haggling and tense hours of brinksmanship.
Substitution	Congress made a final agreement Fri. to avoid government shutdown, after days of haggling and tense hours of brinkmanship.
Deletion	Congressional leaders reached agreement Friday to avert a shutdown of federal government, after haggling and tense hours.

Fig 1: Substitution-based vs. Deletion-based Compression for a Congressional Tweet.[2]

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. They extend the previous approach of paraphrase probability rankings by also utilizing a monolingual distributional similarity, which uses a method described by Van Durme and Lall (2010)[6] to derive approximate cosine similarity scores over feature counts using single token and independent left-and-right contexts. 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. used 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 performed poorly under all cosine-similarity thresholds between 0.65 and 0.95 at a 0.10 increment. They posited 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.

III Method Overview

The process of paraphrastic sentence compression involves a number of steps and tools to reach the final product. Many of these steps rely on previous work done in the research domain of machine translation.

Word Alignment:

In order to prepare parallel corpora for phrase extraction, word alignments must be formed from parallel sentence data. Popular corpora such as Europarl, News Commentary, and the Bible all provide parallelized sentence alignments between English and a variety of foreign languages, which must be tokenized and normalized to deter data splitting from features such as irregular capitalization. This can be done using the Python NLTK and Punkt tokenizer models, as well as a simple Perl script for normalization.

Word alignment can be performed using an unsupervised EM-based approach on unannotated parallel corpora. The experiments presented in this paper utilize IBM Model 1 and an iterative HMM run using the Berkeley aligner to accomplish this task. IBM Model 1 is a simple and effective model, but assumes that all alignments are equally likely and does not take word locality into account. The HMM model provides a word alignment refinement by taking locality into consideration when calculating translation probabilities, making longer jumps when switching from translating one word to another less likely.

Phrase Alignment:

The basic phrase alignment technique follows that of phrase-based statistical machine translation introduced in Koehn et al (2003). In particular, their technique for aligning phrases involves incrementally building longer phrases from words and phrases with adjacent alignment points. Candidates are then ranked using a two-step process: they 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. The latter metric will later be used for the evaluation of the new, proposed phrase acquisition techniques.

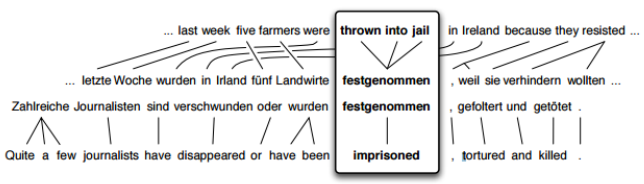

Fig 2: Phrase Alignment and basic paraphrase extraction using a bilingual parallel corpus.

Extracting Phrases from Multiple Bilingual Corpora:

There is nothing limiting the paraphrase extraction method from drawing on candidates in different target languages, provided the corpus has been parallelized in a multitude of different languages. By drawing on paraphrases from multiple parallel corpora, the size of the training corpus increases dramatically, and the system is able to draw on the syntactic, semantic, and cultural intricacies of different languages to connect different English phrases.

Extracting Cross-Domain Phrases using Deep-Linking Alignments:

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.

Deep-linking alignments is simple once phrase alignments are calculated. Phrase alignments will produce a large mapping from English words and phrases to other foreign phrases. A short deep-link would utilize two parallel corpora with the same language pairings; however, multiple parallel corpora can be chained together using corpora that does not involve the source language. A number of different configurations can be applied to adjust the speed and extensiveness of phrase acquisition through deeply-linked alignments, including depth limits, language variety, and language orderings.

Optimizing Language Link Order:

As noted, there are many factors that come into play with deeply-linked alignments; in particular, language origins and their orders with the system can play a large role in the success of deep phrase 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 phrase alignments are therefore reliant on the compression and decompression of their semantic expressions as they are converted from one language to the next. Language re-orderings, e.g. from ENG-GER-SPN-ENG to ENG-SPN-GER-ENG, can have vast effects on the quantity of paraphrases found through deep-linking. Manual optimization is briefly experimented with using the Europarl and Bible corpora.

IV Experimental Evaluation

The experiments involving paraphrases extracted from multiple disjoint bilingual corpora were evaluated using the Europarl de-en and fr-en corpora, as well as the News Commentary de-en corpus. Cross-domain linking was evaluated with a depth of 1 on Europarl de-en and News Commentary de-en, and with a depth of 2 on Europarl en-de, Bible de-sp, and Bible sp-en, where the depth level indicates the number of auxiliary corpora used from translation of the source phrase to the paraphrase. Another test was conducted on Europarl en-sp, Bible sp-de, and Bible de-en. The Europarl parallel corpora were shortened to 300,000 sentences due to computational time and memory constraints -- the word alignment phase had trouble completing, even with 300,000 sentences on 2-4 gigabytes of memory on Condor.

In evaluating newly-acquired paraphrases, the monolingual distributional similarity scores used by Napoles et al. and Van Durme and Lall were calculated. Although previous papers have used web-scale Google n-gram corpora for training, I did not have enough time or memory space to build a full semantic space model using the 5-gram models. Instead, I used a pre-built WordNet-based distributional similarity model. The scores provided using this model were compared with the paraphrase scores from those acquired using the basic paraphrase acquisition method described by Napoles et al. An evaluation set of 50 common initial English phrases were used.

under control	**green light**	**knock back**
sooner or later	**great care**	**study in detail**
military force	**crystal clear**	**claimed responsibility for**
long ago	**at work**	**in check**

Fig 3: An example subset of the common phrases used for paraphrase evaluation.

Corpora	Max	Avg	Num
Europarl de-en	0.81	0.43	6.41
Europarl fr-en	0.82	0.42	6.32
News Commentary de-en	0.80	0.67	2.70
Europarl de-en, fr-en (in parallel)	0.82	0.43	6.80
Europarl en-de, NC de-en	0.55	0.38	+0.52
Euro en-de, Bible de-sp, Bible sp-en	0.46	0.29	+0.68
Euro en-sp, Bible sp-de, Bible de-en	0.44	0.31	+0.72

Fig 4: (Average) maximum + per-phrase monolingual score and average acquired paraphrase count for basic, parallelized, and deep-linked corpora.

Multiple Corpora in Parallel:

From figure 4, it is easy to see that acquiring paraphrases from multiple parallel corpora, even from the same dataset, can yield more phrases and a higher overall accuracy. The method can gather direct paraphrases from both the English-German and English-French maps, acquiring the top paraphrases from both corpora and improving the overall Max+Avg monolingual score. However, it is important to note that this is only a result of using the same corpus in different parallel languages; using different corpora, such as News Commentary de-en and Europarl de-en, will maximize the average maximization score, but the average scores would likely be diluted.

Cross-Domain Corpora in Series:

On the other hand, deep-linking phrase alignments across varied domain corpora produce less satisfying results. Europarl +NC produce an average of 0.52 more paraphrases per phrase, and Europarl+Bible produce approximately 0.70 more at a depth of 2 over the basic single-corpus acquisition method described by Napoles et al. While these numbers by themselves are not problematic, both the average per-phrase monolingual scores and the total average maximum score are significantly lower than the initial paraphrases found by the basic method -- an average of 0.13 less than the basic Europarl per-phrase score, and an average of 0.36 less than the average maximum. From the experiments run, it is apparent that while deep-linking may find new paraphrases, they are not likely to be of good quality, based on the monolingual cosine similarity metric.

At the same time, the experimental corpora used leaves room for debate. The News Commentary and Bible corpora are significantly smaller than Europarl at 21,185 and 62,206 sentences. This would likely have functioned as a bottleneck for translational variation, as there is a severe drop in the amount of data that can be used to find additional paraphrases. In addition, the Bible corpus is written in almost an entirely different style from modern English, and its effect can be seen from the results: the method is able to find approximately 0.7 more paraphrases per test phrase, but the quality of these paraphrases are significantly lower than the ones found from the Europarl-News Commentary links.

From the initial evaluations, it is clear that more in-depth testing should be done to validate the results. Larger and more modern corpora should be used, as both NC and Bible were an order of magnitude smaller than 300,000 sentences, and the Bible uses vastly different wordings from modern English literature.

V Future Work

As noted in Napoles et al. (2003), there are detrimental issues with their chosen method of candidate compression ranking due to erroneous paraphrase substitution using phrases with the same syntactic category and distributional similarity, but different semantics. More sophisticated techniques for compression rankings can utilize optimization procedures over a variety of features beyond minimal length and monolingual similarity, including more diverse syntactic and semantic features, and various word-sense disambiguation methods for filtering out unlikely paraphrases.

More work can be done with regards to multilingual parallel corpora, including statistical techniques for deep-linking of multiple parallel corpora. Research on optimization of the implicit compression and decompression of semantic phrases between languages can reveal vast improvements in the number of paraphrase candidates produced for sentence compression tools.

In regards to correlations between corpus size and paraphrase acquisition, more research can be done on analyzing the curve of paraphrase quantity and quality gains as the corpus size increases, both as a result of increased depth limits, and as a result of larger corpora. Different types of phrases may also effect the quantity and quality of found paraphrases, and most research so far has evaluated results on less than 300 phrases or sentences.

There is also 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 lingual analyses of social media, as the vast assortment of tools traditionally used for linguistic analysis would not need to be modified to fit the domain. Such decompression techniques would likely focus more heavily on reliable paraphrase ranking methods, as character length no longer becomes a constraining feature.

In the future, corpora containing a rich amount of abbreviated and paraphrased text can provide a much more contextualized and naturalistic sentence compression model, as these paraphrases are often character-based compressions formed manually by users themselves. Recent parallel corpora created for microblogging websites like Twitter and Weibo have emerged in the past two years, but due to 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.

VI Conclusion

Information is becoming more mobile and accessible for a large number of digital consumers as technology becomes cheaper and more entwined with everyday life. Sentence compression plays a major role in improving the mobility and readability of information for use in microblogging websites, subtitling, data size compression, and a wide variety of other fields. This paper explored two new paraphrase acquisition techniques for use in substitution-based compression, combining multiple bilingual parallel corpora from varied domains in parallel and in series to expand the search for viable paraphrasal options. The experimental results show promise in using noisy methods of paraphrase extraction to acquire new paraphrases. The results show that the new methods find paraphrases of relatively poor quality, but the experimental flaws revealed during evaluation indicate that the assumptions made during cross-domain linking and deep-linking phrase alignments can still provide new viable options for paraphrasing. In the future, evaluation with larger and more varied corpora, along with more sophisticated phrase alignment and ranking tools, can reveal more convincing improvements. It is believable that future improvements in paraphrase acquisition and rank optimization can eventually provide concrete evidence of substitution-based compression as a more accurate and reliable method of compression over deletion-based methods.

VII References

[1] Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 703–710. AAAI Press / The MIT Press.

[2] Courtney Napoles, Chris Callison-Burch, Juri Ganitkevitch and Benjamin Van Durme. 2011. Paraphrastic Sentence Compression with a Character-based Metric: Tightening without Deletion. In *Proceedings of Workshop on Monolingual Text-To-Text Generation (Text-To-Text-2011)*.

[3] Barnard, C., and Callison-Burch, C. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.*

[4] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/ NAACL.*

[5] Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of ACL.*

[6] Benjamin Van Durme and Ashwin Lall. 2010. Online generation of locality sensitive hash signatures. In *Proceedings of ACL, Short Papers*.