Evaluating the grammaticality of the mapping between source sentence and target sentence in gold standard corpora for compression

Abstract

We present a semi-automatic error analysis approach that demonstrates a limitation to the current commonly adopted paradigm for sentence compression that arises from the strong assumption of locality of the decision making process in the search for an acceptable derivation. Based on our error analysis, we present promising new directions in statistical compression work.

1 Introduction

In this paper we present a semi-automatic error analysis approach that demonstrates a limitation to the current commonly adopted paradigm for sentence compression (Knight and Marcu, 2000; Turner and Charniak, 2005; McDonald, 2006; Clark and Lapata 2006). In addition to presenting our findings along these lines, we argue for the potential contribution of this semi-automatic error analysis technique for other areas of language technologies such as statistical machine translation (SMT) (Yamada and Knight, 2001) and SMT with paraphrasing (Callison-Burch et. al., 2006).

Specifically for statistical compression, a simplifying assumption is made that compression is accomplished strictly by means of word deletion. Furthermore, each sequence of contiguous words that are dropped from a source sentence is considered independently of other sequences of words dropped from other portions of the sentence, so that the features that predict whether deleting a sequence of words is preferred or not is based on local considerations. A similar assumption of locality is made within much typical work within the areas of SMT and SMT with paraphrasing. This simplistic approach allows all possible derivations to be modeled efficiently within the search space, and for decoding to occur efficiently as well using a dynamic programming algorithm.

In theory, it should be possible to learn how to do compression from a corpus of source-target sentence pairs, given enough examples and sufficiently expressive features. However, our analysis casts doubt that this framework with its strong assumptions of locality is sufficiently powerful to learn the types of example compressions frequently found in corpora of human generated gold standard compressions regardless of how expressive the features are.

Work in sentence compression has been somewhat hampered by the tremendous cost involved in producing a gold standard corpus. Because of this tremendous cost, the same gold standard corpora are used in many different published studies more or less as a black box, without a tremendous amount of scrutiny about the limitations on the learnability of the desired target systems resulting from inconsistencies between subtleties in the process by which humans generate the gold standard compressions from the source sentences and the strong locality assumptions inherent in the frameworks.

Typically, the humans who have participated in the construction of these corpora are instructed to preserve grammaticality and to produce compressions by deletion. Human ratings of the gold standard compressions by separate judges confirm that the human developers have literally followed the instructions, and have produced compressions that are themselves grammatical. Nevertheless, what we demonstrate with our error analysis is that what they haven't consistently done is preserve a grammatical mapping between source and target sentences, which places limitations on how well the patterns of compression can be learned using the current state-of-the-art paradigm.

In the remainder of this paper, we discuss relevant work in sentence compression, statistical machine translation, and paraphrase that is applicable. We then introduce our semi-automatic error analysis technique. Next we discuss the error analysis itself and the conclusions we draw from it. Finally, we conclude with future directions for broader application of this error analysis technique.

2 Related Work

Knight and Marcu (2000) present two approaches to the sentence compression problem- one using a noisy channel model and the other using a decision-based model. Subsequent work (McDonald, 2006) has demonstrated an advantage for a soft constraint approach where a discriminative model learns whether it is advisable to drop all of the words between a pair of words in the source sentence. Features in this system are defined over pairs of words in the compressed sentence, while also using the words present in the sentence that were dropped in order to obtain the compressed sentence. The discriminative learning system can handle features that overlap, and the learner sets the weights of each feature relative to the others so as to optimize the accuracy of the model over the observed data.

We use McDonald (2006) proposed model as a foundation for our work because its soft constraint approach allows for natural integration of a variety of classes of features. In our prior work we have explored the potential for improving the performance of a compression system by including additional, more sophisticated syntactically motivated features than those included in previously published models. In this paper, we evaluate the gold standard corpus itself using similar syntactic grammar policies.

3 Grammar Policy Extraction

In the domain of Sentence Compression, the corpus consists of source sentence and gold standard compressed sentence. The gold standard corpus used in the analysis we present in this paper was constructed by combining the training sets from two commonly used corpora in compression research, namely the Ziff-Davis set (Knight and Marcu, 2002) consisting of 1055 sentences, and a partial Broadcast News Corpus (Clarke and Lapata, 2006) consisting of 1070 sentences. We hypothesize certain grammar policies that intuitively should be followed while deriving the target-compressed sentence from the source sentence if the mapping between source and target sentences is grammatical. These policies, based on the MST (McDonald, 2005) dependency parse structure of the source sentence, are as follows:

1. The syntactic root word of a sentence should be retained in the compressed sentence.
2. If a verb is retained in the compressed sentence, then the dependent subject of that verb should also be retained.
3. If a verb is retained in the compressed sentence, then the dependent object of that verb should also be retained.
4. If the verb is dropped in the compressed sentence then its arguments, namely subject, object, prepositional phrases etc., should also be dropped.
5. If the Preposition in a Prepositional phrase(PP) is retained in the compressed sentence, then the dependent Noun Phrase(NP) of that Preposition should also be retained.
6. If the head noun of a Noun phrase(NP) within a Prepositional phrase is retained in the compressed sentence, then the syntactic parent Preposition of the NP should also be retained.
7. If a Preposition, the syntactic head of a Prepositional phrase(PP) is dropped in the compressed sentence, then the whole PP, including dependent Noun phrase in that PP, should also be dropped.
8. If the head noun of a Noun phrase within a Prepositional phrase(PP) is dropped in the compressed sentence, then the syntactic parent Preposition of the PP should also be dropped.

These grammar policies represent probable phrase structure to be dropped or retained in the compression and are thus similar to the syntactic features in McDonald (2006). But there is a fundamental difference in the way these policies are computed. In McDonald (2006), the features are computed locally over adjacent words y_{i-1} & y_i in the compression and the words dropped from the original sentence between that range. In cases where the syntactic structure of the involved words extend beyond this range, the extracted features are not able to capture the syntactic dependencies,. Whereas in our system, the policies are computed globally over the complete sentence without specifying any range of words. Let us consider the following sentence from the Clark-Lapata Corpus (bold represents dropped words):

1. The$_1$ leaflet$_2$ given$_3$ to$_4$ Labour$_5$ **activists**$_6$ mentions$_7$ none$_8$ of$_9$ these$_{10}$ things$_{11}$**.**

According to Policy 2, since the verb 'mentions' is retained in the global context is taken into account while evaluating the verb 'mentions'. In McDonald 2006, examining the compression, the Subject of the verb 'The leaflet' should also be retained. This policy can only be captured if the global context is taken into account while evaluating the verb 'mentions'. In McDonald (2006), the looking at the range $y_{i-1} = 5$ and $y_i = 7$ for the verb 'mentions', we will not be able to compute whether the subject(1,2) was retained in the compression or not.

Now we can evaluate each sentence in the corpus to determine whether a particular policy was applicable and if applicable then whether it was violated. Table 1 shows the summary of the evaluation of all the sentences in the two corpuses.

4 Evaluation

In this section we discuss the results from evaluating the 8 grammar policies discussed in Section 3 over two commonly used training corpora for statistical compression work, namely the 1055 sentence Ziff-Davis corpus and the 1070 sentence Clark-Lapata corpus. The striking finding is that for every one of the policies discussed in the previous section, they are broken for at least 10% of the sentences where they apply in the training corpus, and sometimes as much as 72% of the times when they apply. For most policies, the proportion of sentences where the policy is broken when applied is a minority of cases. Thus, based on this, we can expect that grammar oriented features derived from a syntactic analysis of the source and/or target sentences in the gold standard could be used to improve the performance of compression systems that don't make use of syntactic information to that extent. However, the sizeable percentage of time when these very intuitive grammar policies are broken when applied indicates that there is a limited extent to which this type of feature is likely to contribute to improved performance.

One observation we make from Table 1 is that while the proportion of sentences where the policies apply as well as the proportion of sentences where the policies are broken when applied are highly correlated between the two corpora, it is not identical. Thus, again, while we predict that using dependency syntax features might improve performance of compression systems within a single corpus, we would expect degradation between corpora resulting from the differences in the extent to which these grammar inspired policies are kept.

	Ziff-Davis (% Applied)	Ziff-Davis (% Broken when Applied)	Clark-Lapata (% Applied)	Clark-Lapata (% Broken when Applied)
Policy1	100%	34%	100%	14%
Policy2	66%	18%	84%	18%
Policy3	50%	10%	61%	24%
Policy4	59%	59%	46%	72%
Policy5	62%	17%	77%	27%
Policy6	65%	22%	79%	29%
Policy7	57%	25%	58%	40%
Policy8	55%	16%	58%	36%

Table 1: Summary of evaluation of grammar policies over the Ziff-Davis training set and Clark-Lapata training set.

Beyond the above evaluation illustrating the extent to which grammar inspired policies are broken in human generated gold standard corpora, interesting insights into potential new directions for work on statistical compression can be obtained by taking a close look at typical examples from the Clark-Lapata corpus where the policies are broken in the gold standard corpora.

1. The attempt to **put flesh and blood on the skeleton** structure **of** a **possible** united Europe emerged.
2. Annely **has used the gallery** 's three floors **to** divide the exhibits into three **distinct** groups.
3. Labour **has said it** will scrap the system.

In Sentence 1, retaining the dependent Noun 'structure' of the dropped Preposition 'on' in the PP breaks Policy 7. Such Noun Phrase to Infinitive Phrase transformation changes the syntactic structure of the sentence. Sentence 2 also breaks several policies namely – Policy 1, 4 and 7. The syntactic root 'has' of the sentence is dropped. Also the main verb 'has used' is dropped while retaining the

Subject 'Annely' of the sentence. Breaking Policies 1, 2 and 4, the human annotator replaced the pronoun 'it' to the noun 'Labour' the subject of a dropped verb 'has said'. Such anaphora resolution cannot be done without relevant context, which is usually not available in the domain of Sentence Compression. Such varied transformations, made by human annotators, in the syntactic structure of the sentence are against intuition making them very hard to be captured by the syntactic features in current compression systems.

5 Conclusions and Current Directions

In this paper we have introduced a semi-automatic error analysis technique that was used to investigate potential impact and limitations of adding dependency parse features to the problem of statistical compression. We have argued that the reason for the limitation arises from the strong assumption of the local nature of the decisions that are made in obtaining the system-generated compression from a source sentence.

Based on our error analysis, we have discussed promising new directions in statistical compression work. Because other related technologies such as statistical machine translation and statistical paraphrase are based on similar paradigms with very similar strong assumptions of the local nature of decisions that are made in the search for an acceptable derivation, we argue both that it is likely that the same issues related to the construction of the gold standard corpus likely apply and that a similar semi-automatic error analysis approach could be applied in order to assess the extent to which this is true. In our ongoing work we plan to conduct a similar error analysis for these problems in order to evaluate the generality of the findings reported here. Furthermore, we plan to implement and evaluate the proposed extensions to typical statistical compressions systems to evaluate their impact on the learnability of the gold standard corpora used for training.

Acknowledgments

This work was funded in part by XXXX grant number XXXX.

References

Callison-Burch, Chris, Philipp Koehn, and Miles Osborne. 2006. *Improved statistical machine translation using paraphrases.* In Proceedings of the Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics, pages 17–24, New York, NY.

James Clarke and Mirella Lapata. 2006. *Constraint-Based Sentence Compression: An Integer Programming Approach.* Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions (ACL-2006), pages 144-151, 2006.

James Clarke and Mirella Lapata. 2006. *Models for Sentence Compression: A Comparison across Domains, Training Requirements and Evaluation Measures. In* Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 377-384. Sydney, Australia.

Kevin Knight and Daniel Marcu. 2000. *Statistics-Based Summarization – Step One: Sentence Compression.* Proceedings of AAAI-2000, Austin, TX, USA.

Knight, Kevin and Daniel Marcu. 2002. *Summarization beyond sentence extraction: a probabilistic approach to sentence compression*. Artificial Intelligence 139(1):91–107.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. *Online large-margin training of dependency parsers.* Proc. ACL.

Ryan Mcdonald, 2006. *Discriminative sentence compression with soft syntactic constraints*. Proceedings of the 11th EACL. Trento, Italy, pages 297--304.

J. Turner and E. Charniak. 2005. *Supervised and unsupervised learning for sentence compression*. In Proc. ACL.

Yamada, K., & Knight, K. (2001). *A syntax-based statical translation model.* In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pp. 523–530, Toulouse, France.