Evaluating the Syntactic Transformations in Gold Standard Corpora for Statistical Sentence Compression

Naman K. Gupta, Sourish Chaudhuri, Carolyn P. Rosé
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{nkgupta,sourishc,cprose}@cs.cmu.edu

Abstract

We present a policy-based error analysis approach that demonstrates a limitation to the current commonly adopted paradigm for sentence compression. We demonstrate that these limitations arise from the strong assumption of locality of the decision making process in the search for an acceptable derivation in this paradigm.

1 Introduction

In this paper we present a policy-based 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).

Specifically, in typical statistical compression approaches, 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 solely on local considerations. This simplistic approach allows all possible derivations to be modeled and decoded efficiently within the search space, using a dynamic programming algorithm.

In theory, it should be possible to learn how to generate effective compressions using 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 almost as a black box. This is done with little scrutiny of the limitations on the learnability of the desired target systems. These limitations result from inconsistencies due to the subtleties in the process by which humans generate the gold standard compressions from the source sentences, and from the strong locality assumptions inherent in the frameworks.

Typically, the humans who have participated in the construction of these corpora have been 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 largely grammatical. Nevertheless, what we demonstrate with our error analysis is that they have used meaning preserving transformation that didn't consistently preserve the grammatical relations from the source sentence while transforming source sentences into target sentences. This places limitations on how well the preferred patterns of compression can be learned using the current paradigm and existing corpora.

In the remainder of the paper, we discuss relevant work in sentence compression. We then introduce our policy-based error analysis technique. Next we discuss the error analysis itself and the conclusions we draw from it. Finally, we conclude

Proceedings of NAACL HLT 2009: Short Papers, pages 145–148,
Boulder, Colorado, June 2009.

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 to make local decisions about dropping a sequence of words from the source sentence in order to produce the target compression. Features in this system are defined over pairs of words in the source sentence, with the idea that the pair of words would appear adjacent in the resulting compression, with all intervening words dropped. Thus, the features represent this transformation, and the feature weights are meant to indicate whether the transformation is associated with good compressions or not.

We use McDonald's (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, even overlapping 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 sentences each paired with a gold standard compressed sentence. Most of the above related work has been evaluated using the following 2 corpora, namely the Ziff-Davis (ZD) set (Knight and Marcu, 2002) consisting of 1055 sentences, and a partial Broadcast News Corpus (CL Corpus) (Clarke and Lapata, 2006) originally consisting of 1619 sentences, of which we used 1070 as the training set in our development work as well as in the error analysis below. Hence, we use these two popular corpora to present our work. 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 produced via grammatical transformations. The basic idea behind these policies grows out of the same ideas motivating the syntactic features used in McDonald (2006). These policies, extracted using 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 make predictions about where, in the phrase structure, constituents are likely to be dropped or retained in the compression. Thus, these policies have similar motivation to the syntactic features in the McDonald (2006) model. However, there is a fundamental difference in the way these policies are computed. In the McDonald (2006) model, the features are com-

puted locally over adjacent words y_{i-1} & y_i in the compression and the words dropped from the original sentence between that word range y_{i-1} & y_i. In cases where the syntactic structure of the involved words extends beyond this range, the extracted features are not able to capture all of the relevant syntactic dependencies. On the other hand, in our analysis the policies are computed globally over the complete sentence without specifying any range of words. As an illustrative example, let us consider the following sentence from the CL 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, the subject of the verb 'the leaflet' should also be retained. In the McDonald (2006) model, by looking at the local 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. So this policy can be captured only if the global context is taken into account while evaluating the verb 'mentions'.

Now we 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 corpora. Column 2 in the table shows the percentage of sentences in the ZD Corpus where the respective policies were applicable. And column 3 shows the percentage of sentences where the respective policies were violated, whenever applicable. Columns 4 and 5 show respective percentages for the CL corpus.

4 Evaluation

In this section we discuss the results from evaluating the 8 grammar policies discussed in Section 3 over the ZD and CL corpora, as discussed above.

The policies were evaluated with respect to whether they applied in a sentence, i.e., whether the premise of the "if … then" rule is true in the sentence, and whether the policy was broken when applied, i.e., if the premise is true but the consequent is false. The striking finding is that for every one of the policies discussed in the previous section, they are violated for at least 10% of the sentences where they applied, and sometimes as much as 72%. For most policies, the proportion of sentences where the policy is violated when applied is a minority of cases. Thus, based on this, we can expect that grammar oriented features motivated by these policies and 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 noticeable proportion of violations with respect to some of the policies indicate that there is a limited extent to which these types of features can contribute towards improved performance.

One observation we make from Table 1 is that while the proportion of sentences where the policies (Columns 2 and 4) apply as well as the proportion of sentences where the policies are broken when applied (Columns 3 and 5) are highly correlated between the two corpora. Nevertheless, the distributions are not identical. Thus, again, while we predict that using this style of dependency syntax features might improve performance of compression systems within a single corpus, we would not expect trained models that rely on these syntactic dependency features to generalize in an ideal way between corpora.

	ZD (% Applicable)	ZD (% Violations when Applicable)	CL (% Applicable)	CL (% Violations when Applicable)
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 (ZD) training set and Clark-Lapata (CL) training set.

Beyond the above evaluation illustrating the extent to which grammar inspired policies are violated in human generated gold standard corpora, interesting insights into challenges that must be addressed in order to improve performance can be obtained by taking a close look at typical examples from the CL corpus where the policies are broken in the

gold standard corpora (bold represents dropped words).

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. Labor **has said it** will scrap the system.
4. Montenegro 's **sudden** rehabilitation of Nicholas **'s memory** is **a** popular **move**.

In Sentence 1, retaining the dependent Noun *structure* of the dropped Preposition *on* in the PP violates Policy 7. Such a NP to Infinitive Phrase transformation changes the syntactic structure of the sentence. Sentence 2 also breaks several policies, namely Policies 1, 4 and 7. The syntactic root *has* is dropped. Also the main verb *has used* is dropped while retaining the Subject *Annely*. In Sentence 3, breaking Policies 1, 2 and 4, the human annotators replaced the pronoun *it* with the noun *Labor*, the subject of a dropped verb 'has said'. Such anaphora resolution cannot be done without relevant context, which is not available in strictly local paradigms of sentence compression. In Sentence 4, policies 3. 5 and 8 are violated. Transformations like substituting *Nicholas's memory* by the metonym *Nicholas* and *popular move* by *popular* need to be identified and analyzed. Such varied transformations, made in the syntactic structure of the sentences by human annotators, are counter-intuitive, making them hard to be captured in the linear models learned in association with the syntactic features in current compression systems.

5 Conclusions and Current Directions

In this paper we have introduced a policy-based error analysis technique that was used to investigate the potential impact and limitations of adding a particular style of dependency parse features to typical statistical compression systems. 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.

Other related technologies such as statistical machine translation and statistical paraphrase are based on similar paradigms with similar assumptions of the local nature of decisions that are made in the search for an acceptable derivation. We conjecture both that it is likely that the same issues related to the construction of the gold standard corpora likely apply and that a similar policy-based error analysis approach could be used in order to assess the extent to which this is true and identify possible directions for improving performance. 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.

Acknowledgments

This work was funded in part by the Office of Naval Research grant number N00014510043.

References

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.* 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.

Jenine Turner and Eugene Charniak. 2005. *Supervised and unsupervised learning for sentence compression*. Proc. ACL.