A Two-step Approach to Sentence Compression of Spoken Utterances

Dong Wang, Xian Qian, Yang Liu
The University of Texas at Dallas
dongwang,qx,yangl@hlt.utdallas.edu

Abstract

This paper presents a two-step approach to compress spontaneous spoken utterances. In the first step, we use a sequence labeling method to determine if a word in the utterance can be removed, and generate n-best compressed sentences. In the second step, we use a discriminative training approach to capture sentence level global information from the candidates and rerank them. For evaluation, we compare our system output with multiple human references. Our results show that the new features we introduced in the first compression step improve performance upon the previous work on the same data set, and reranking is able to yield additional gain, especially when training is performed to take into account multiple references.

1 Introduction

Sentence compression aims to preserve the most important information in the original sentence with fewer words. It can be used for abstractive summarization where extracted important sentences often need to be compressed and merged. For summarization of spontaneous speech, sentence compression is especially important, since unlike fluent and well-structured written text, spontaneous speech contains a lot of disfluencies and much redundancy. The following shows an example of a pair of source and compressed spoken sentences[1] from human annotation (removed words shown in bold):

[**original sentence**]

and then **um** *in terms of the source* **the things uh** *the only things that we had on there* **I believe** *were whether...*

[**compressed sentence**]

and then in terms of the source the only things that we had on there were whether...

In this study we investigate sentence compression of spoken utterances in order to remove redundant or unnecessary words while trying to preserve the information in the original sentence. Sentence compression has been studied from formal text domain to speech domain. In text domain, (Knight and Marcu, 2000) applies noisy-channel model and decision tree approaches on this problem. (Galley and Mckeown, 2007) proposes to use a synchronous context-free grammars (SCFG) based method to compress the sentence. (Cohn and Lapata, 2008) expands the operation set by including insertion, substitution and reordering, and incorporates grammar rules. In speech domain, (Clarke and Lapata, 2008) investigates sentence compression in broadcast news using an integer linear programming approach. There is only a few existing work in spontaneous speech domains. (Liu and Liu, 2010) modeled it as a sequence labeling problem using conditional random fields model. (Liu and Liu, 2009) compared the effect of different compression methods on a meeting summarization task, but did not evaluate sentence compression itself.

We propose to use a two-step approach in this paper for sentence compression of spontaneous speech utterances. The contributions of our work are:

- Our proposed two-step approach allows us to incorporate features from local and global levels. In the first step, we adopt a similar sequence labeling method as used in (Liu and Liu, 2010), but expanded the feature set, which

[1] For speech domains, "sentences" are not clearly defined. We use sentences and utterances interchangeably when there is no ambiguity.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 166–170,
Jeju, Republic of Korea, 8-14 July 2012.

results in better performance. In the second step, we use discriminative reranking to incorporate global information about the compressed sentence candidates, which cannot be accomplished by word level labeling.

- We evaluate our methods using different metrics including word-level accuracy and F1-measure by comparing to one reference compression, and BLEU scores comparing with multiple references. We also demonstrate that training in the reranking module can be tailed to the evaluation metrics to optimize system performance.

2 Corpus

We use the same corpus as (Liu and Liu, 2010) where they annotated 2,860 summary sentences in 26 meetings from the ICSI meeting corpus (Murray et al., 2005). In their annotation procedure, filled pauses such as "uh/um" and incomplete words are removed before annotation. In the first step, 8 annotators were asked to select words to be removed to compress the sentences. In the second step, 6 annotators (different from the first step) were asked to pick the best one from the 8 compressions from the previous step. Therefore for each sentence, we have 8 human compressions, as well a best one selected by the majority of the 6 annotators in the second step. The compression ratio of the best human reference is 63.64%.

In the first step of our sentence compression approach (described below), for model training we need the reference labels for each word, which represents whether it is preserved or deleted in the compressed sentence. In (Liu and Liu, 2010), they used the labels from the annotators directly. In this work, we use a different way. For each sentence, we still use the best compression as the gold standard, but we realign the pair of the source sentence and the compressed sentence, instead of using the labels provided by annotators. This is because when there are repeated words, annotators sometimes randomly pick removed ones. However, we want to keep the patterns consistent for model training – we always label the last appearance of the repeated words as 'preserved', and the earlier ones as 'deleted'. Another difference in our processing of the corpus from the previous work is that when aligning the original and the compressed sentence, we keep filled pauses and incomplete words since they tend to appear together with disfluencies and thus provide useful information for compression.

3 Sentence Compression Approach

Our compression approach has two steps: in the first step, we use Conditional Random Fields (CRFs) to model this problem as a sequence labeling task, where the label indicates whether the word should be removed or not. We select n-best candidates ($n = 25$ in our work) from this step. In the second step we use discriminative training based on a maximum Entropy model to rerank the candidate compressions, in order to select the best one based on the quality of the whole candidate sentence, which cannot be performed in the first step.

3.1 Generate N-best Candidates

In the first step, we cast sentence compression as a sequence labeling problem. Considering that in many cases phrases instead of single words are deleted, we adopt the 'BIO' labeling scheme, similar to the name entity recognition task: "B" indicates the first word of the removed fragment, "I" represents inside the removed fragment (except the first word), and "O" means outside the removed fragment, i.e., words remaining in the compressed sentence. Each sentence with n words can be viewed as a word sequence $X_1, X_2, ..., X_n$, and our task is to find the best label sequence $Y_1, Y_2, ..., Y_n$ where Y_i is one of the three labels. Similar to (Liu and Liu, 2010), for sequence labeling we use linear-chain first-order CRFs. These models define the conditional probability of each labeling sequence given the word sequence as:

$$p(Y|X) \propto \\ exp \sum_{k=1}^{n} (\sum_j \lambda_j f_j(y_k, y_{k-1}, X) + \sum_i \mu_i g_i(x_k, y_k, X))$$

where f_j are transition feature functions (here first-order Markov independence assumption is used); g_i are observation feature functions; λ_j and μ_i are their corresponding weights. To train the model for this step, we use the best reference compression to obtain the reference labels (as described in Section 2).

In the CRF compression model, each word is represented by a feature vector. We incorporate most of the features used in (Liu and Liu, 2010), including unigram, position, length of utterance, part-of-speech tag as well as syntactic parse tree tags. We did not use the discourse parsing tree based features because we found they are not useful in our experiments. In this work, we further expand the feature set in order to represent the characteristics of disfluencies in spontaneous speech as well as model the adjacent output labels. The additional features we introduced are:

- the distance to the next same word and the next same POS tag.
- a binary feature to indicate if there is a filled pause or incomplete word in the following 4-word window. We add this feature since filled pauses or incomplete words often appear after disfluent words.
- the combination of word/POS tag and its position in the sentence.
- language model probabilities: the bigram probability of the current word given the previous one, and followed by the next word, and their product. These probabilities are obtained from the Google Web 1T 5-gram.
- transition features: a combination of the current output label and the previous one, together with some observation features such as the unigram and bigrams of word or POS tag.

3.2 Discriminative Reranking

Although CRFs is able to model the dependency of adjacent labels, it does not measure the quality of the whole sentence. In this work, we propose to use discriminative training to rerank the candidates generated in the first step. Reranking has been used in many tasks to find better global solutions, such as machine translation (Wang et al., 2007), parsing (Charniak and Johnson, 2005), and disfluency detection (Zwarts and Johnson, 2011). We use a maximum Entropy reranker to learn distributions over a set of candidates such that the probability of the best compression is maximized. The conditional probability of output y given observation x in the maximum entropy model is defined as:

$$p(y|x) = \frac{1}{Z(x)} exp\left[\sum_{i=1}^{k} \lambda_i f(x,y)\right]$$

where $f(x,y)$ are feature functions and λ_i are their weighting parameters; $Z(x)$ is the normalization factor.

In this reranking model, every compression candidate is represented by the following features:

- All the bigrams and trigrams of words and POS tags in the candidate sentence.
- Bigrams and trigrams of words and POS tags in the original sentence in combination with their binary labels in the candidate sentence (delete the word or not). For example, if the original sentence is "so I should go", and the candidate compression sentence is "I should go", then "so_I_10", "so_I_should_100" are included in the features (1 means the word is deleted).
- The log likelihood of the candidate sentence based on the language model.
- The absolute difference of the compression ratio of the candidate sentence with that of the first ranked candidate. This is because we try to avoid a very large or small compression ratio, and the first candidate is generally a good candidate with reasonable length.
- The probability of the label sequence of the candidate sentence given by the first step CRFs.
- The rank of the candidate sentence in 25 best list.

For discriminative training using the n-best candidates, we need to identify the best candidate from the n-best list, which can be either the reference compression (if it exists on the list), or the most similar candidate to the reference. Since we have 8 human compressions and also want to evaluate system performance using all of them (see experiments later), we try to use multiple references in this reranking step. In order to use the same training objective (maximize the score for the single best among all the instances), for the 25-best list, if m reference compressions exist, we split the list into m groups, each of which is a new sample containing one reference as positive and several negative candidates. If no reference compression appears in 25-best list, we just keep the entire list and label the instance that is most similar to the best reference compression as positive.

4 Experiments

We perform a cross-validation evaluation where one meeting is used for testing and the rest of them are used as the training set. When evaluating the system performance, we do not consider filled pauses and incomplete words since they can be easily identified and removed. We use two different performance metrics in this study.

- Word-level accuracy and F1 score based on the minor class (removed words). This was used in (Liu and Liu, 2010). These measures are obtained by comparing with the best compression. In evaluation we map the result using 'BIO' labels from the first-step compression to binary labels that indicate a word is removed or not.

- BLEU score. BLEU is a widely used metric in evaluating machine translation systems that often use multiple references. Since there is a great variation in human compression results, and we have 8 reference compressions, we explore using BLEU for our sentence compression task. BLEU is calculated based on the precision of n-grams. In our experiments we use up to 4-grams.

Table 1 shows the averaged scores of the cross validation evaluation using the above metrics for several methods. Also shown in the table is the compression ratio of the system output. For "reference", we randomly choose one compression from 8 references, and use the rest of them as references in calculating the BLEU score. This represents human performance. The row "basic features" shows the result of using all features in (Liu and Liu, 2010) except discourse parsing tree based features, and using binary labels (removed or not). The next row uses this same basic feature set and "BIO" labels. Row "expanded features" shows the result of our expanded feature set using "BIO" label set from the first step of compression. The last two rows show the results after reranking, trained using one best reference or 8 reference compressions, respectively.

	accuracy	F1	BLEU	ratio (%)
reference	81.96	69.73	95.36	76.78
basic features (Liu and Liu, 2010)	76.44	62.11	91.08	73.49
basic features, BIO	77.10	63.34	91.41	73.22
expanded features	**79.28**	67.37	92.70	72.17
reranking train w/ 1 ref	79.01	**67.74**	91.90	70.60
reranking train w/ 8 refs	78.78	63.76	**94.21**	77.15

Table 1: Compression results using different systems.

Our result using the basic feature set is similar to that in (Liu and Liu, 2010) (their accuracy is 76.27% when compression ratio is 0.7), though the experimental setups are different: they used 6 meetings as the test set while we performed cross validation. Using the "BIO" label set instead of binary labels has marginal improvement for the three scores. From the table, we can see that our expanded feature set is able to significantly improve the result, suggesting the effectiveness of the new introduced features.

Regarding the two training settings in reranking, we find that there is no gain from reranking when using only one best compression, however, training with multiple references improves BLEU scores. This indicates the discriminative training used in maximum entropy reranking is consistent with the performance metrics. Another reason for the performance gain for this condition is that there is less data imbalance in model training (since we split the n-best list, each containing fewer negative examples). We also notice that the compression ratio after reranking is more similar to the reference. As suggested in (Napoles et al., 2011), it is not appropriate to compare compression systems with different compression ratios, especially when considering grammars and meanings. Therefore for the compression system without reranking, we generated results with the same compression ratio (77.15%), and found that using reranking still outperforms this result, 1.19% higher in BLEU score.

For an analysis, we check how often our system output contains reference compressions based on the 8 references. We found that 50.8% of system generated compressions appear in the 8 references when using CRF output with a compression ration of 77.15%; and after reranking this number increases to 54.8%. This is still far from the oracle result – for 84.7% of sentences, the 25-best list contains one or more reference sentences, that is, there is still much room for improvement in the reranking process. The results above also show that the token level measures by comparing to one best reference do not always correlate well with BLEU scores obtained by comparing with multiple references, which shows the need of considering multiple metrics.

5 Conclusion

This paper presents a 2-step approach for sentence compression: we first generate an n-best list for each source sentence using a sequence labeling method, then rerank the n-best candidates to select the best one based on the quality of the whole candidate sentence using discriminative training. We evaluate the system performance using different metrics. Our results show that our expanded feature set improves the performance across multiple metrics, and reranking is able to improve the BLEU score. In future work, we will incorporate more syntactic information in the model to better evaluate sentence quality. We also plan to perform a human evaluation for the compressed sentences, and use sentence compression in summarization.

6 Acknowledgment

This work is partly supported by DARPA under Contract No. HR0011-12-C-0016 and NSF No. 0845484. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or NSF.

References

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 173–180, Stroudsburg, PA, USA. Proceedings of ACL.

James Clarke and Mirella Lapata. 2008. *Global inference for sentence compression an integer linear programming approach*. *Journal of Artificial Intelligence Research*, 31:399–429, March.

Trevor Cohn and Mirella Lapata. 2008. *Sentence compression beyond word deletion*. In *Proceedings of COLING*.

Michel Galley and Kathleen R. Mckeown. 2007. *Lexicalized Markov grammars for sentence compression*. In *Proceedings of HLT-NAACL*.

Kevin Knight and Daniel Marcu. 2000. *Statistics-based summarization-step one: Sentence compression*. In *Proceedings of AAAI*.

Fei Liu and Yang Liu. 2009. *From extractive to abstractive meeting summaries: can it be done by sentence compression?* In *Proceedings of the ACL-IJCNLP*.

Fei Liu and Yang Liu. 2010. *Using spoken utterance compression for meeting summarization: a pilot study*. In *Proceedings of SLT*.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. *Extractive summarization of meeting recordings*. In *Proceedings of EUROSPEECH*.

Courtney Napoles, Benjamin Van Durme, and Chris Callison-Burch. 2011. *Evaluating Sentence Compression: Pitfalls and Suggested Remedies*. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 91–97, Portland, Oregon, June. Association for Computational Linguistics.

Wen Wang, A. Stolcke, and Jing Zheng. 2007. *Reranking machine translation hypotheses with structured and web-based language models*. In *Proceedings of IEEE Workshop on Speech Recognition and Understanding*, pages 159–164, Kyoto.

Simon Zwarts and Mark Johnson. 2011. *The impact of language models and loss functions on repair disfluency detection*. In *Proceedings of ACL*.