Joint Decoding of Tree Transduction Models for Sentence Compression

Jin-ge Yao **Xiaojun Wan** **Jianguo Xiao**
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
Key Laboratory of Computational Linguistic (Peking University), MOE, China
{yaojinge, wanxiaojun, xiaojianguo}@pku.edu.cn

Abstract

In this paper, we provide a new method for decoding tree transduction based sentence compression models augmented with language model scores, by jointly decoding two components. In our proposed solution, rich local discriminative features can be easily integrated without increasing computational complexity. Utilizing an unobvious fact that the resulted two components can be independently decoded, we conduct efficient joint decoding based on dual decomposition. Experimental results show that our method outperforms traditional beam search decoding and achieves the state-of-the-art performance.

1 Introduction

Sentence compression is the task of generating a grammatical and shorter summary for a long sentence while preserving its most important information. One specific instantiation is deletion-based compression, namely generating a compression by dropping words. Various approaches have been proposed to challenge the task of deletion-based compression. Earlier pioneering works (Knight and Marcu, 2000) considered several insightful approaches, including noisy-channel based generative models and discriminative decision tree models. Structured discriminative compression models (McDonald, 2006) are capable of integrating rich features and have been proved effective for this task. Another powerful paradigm for sentence compression should be mentioned here is constraints-based compression,including integer linear programming solutions (Clarke and Lapata, 2008) and first-order Markov logic networks (Huang et al., 2012; Yoshikawa et al., 2012).

A notable class of methods that explicitly deal with syntactic structures are tree transduction models (Cohn and Lapata, 2007; Cohn and Lapata, 2009). In such models a synchronous grammar is extracted from a corpus of parallel syntax trees with leaves aligned. Compressions are generated from the grammar with learned weights. Previous works have noticed that local coherence is usually needed by introducing ngram language model scores, which will make accurate decoding intractable. Traditional approaches conduct beam search to find approximate solutions (Cohn and Lapata, 2009).

In this paper we propose a joint decoding strategy to challenge this decoding task. We address the problem as jointly decoding a simple tree transduction model that only considers rule weights and an ngram compression model. Although either part can be independently solved by dynamic programming, the naive way to integrate two groups of partial scores into a huge dynamic programming chart table is computationally impractical. We provide an effective dual decomposition solution that utilizes the efficient decoding of both parts. By integrating rich structured features that cannot be efficiently involved in normal formulation, results get significantly improved.

2 Motivation

Under the tree transduction models, the sentence compression task is formulated as learning a mapping from an input source syntax tree to a target tree with reduced number of leaves. This mapping is known as a synchronous grammar. The synchronous grammar discussed through out this paper will be synchronous tree substitution grammar (STSG), as in previous studies.

In such formulations, sentence compression is finding the best derivation from a syntax tree that produces a simpler target tree, under the current definition of grammar and learned parameters. Each derivation is attached with a score. For the sake of efficient decoding, the score often decom-

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1828–1833,
October 25-29, 2014, Doha, Qatar.

poses with rules involved in the derivation. A typical score definition for a derivation $\mathbf{y}$ of source tree $\mathbf{x}$ is in such form (Cohn and Lapata, 2008; Cohn and Lapata, 2009):

$$S(\mathbf{x},\mathbf{y}) = \sum_{r \in \mathbf{y}} \mathbf{w}^T \phi_r(\mathbf{x}) + \log P(ngram(\mathbf{y})) \quad (1)$$

The first term is a weighted sum of features $\phi_r(\mathbf{x})$ defined on each rule r. It is plausible to introduce local scores from ngram models. The second term in the above score definition is added with such purpose.

Cohn and Lapata (2009) explained that exact decoding of Equation 1 is intractable. They proposed a beam search decoding strategy coupled with cube-pruning heuristic (Chiang, 2007), which can further improve decoding efficiency at the cost of largely losing exactness in log probability calculations. For efficiency reasons, rich local ngram features have not been introduced as well.

3 Components of Joint Decoding

The score in Equation 1 consists of two parts: sum of weighted rule features and local ngram scores retrieved from a language model. There is an implicit fact that either part can be used alone with slight modifications to generate a coarse candidate compression. Therefore, we can build a joint decoding system that consists of these two independently decodable components.

In this section we will refer to these two independent models as the pure tree transduction model and the pure ngram compression model, described in Section 3.1 and Section 3.2 respectively. There is a direct generalization of the ngram model by introducing rich local features, which results in the structured discriminative models (Section 3.3).

3.1 Pure Tree Transduction model

By merely considering scores from tree transduction rules, i.e. the first part of Equation 1, we can have our scores factorized with rules. Then finding the best derivation from a STSG grammar can be easily solved by a dynamic programming process described by Cohn and Lapata (2007).

This simplified pure tree transduction model can still produce decent compressions if the rule weights are properly learned during training.

3.2 Pure Ngram based Compression

The pure ngram based model will try to find the most locally smooth compression, reflected by having the maximum log probability score of ngrams.

To avoid the trivial solution of deleting all words, we find the target compression with specified length by dynamic programming.

Furthermore, we can integrate features other than log probabilities. This is equivalent to using a structured discriminative model with rich features on ngrams of candidate compressions.

3.3 Structured Discriminative Model

The structured discriminative model proposed by McDonald (2006) defines rich features on bigrams of possible compressions. The score is defined as weighted linear combination of those features:

$$f(\mathbf{x},\mathbf{z}) = \sum_{j=2}^{|\mathbf{z}|} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, L(\mathbf{z}_{j-1}), L(\mathbf{z}_j)) \quad (2)$$

where the function $L(\mathbf{z}_k)$ maps a token $\mathbf{z}_k$ in compression $\mathbf{z}$ back to the index of the original sentence $\mathbf{x}$. Decoding can still be efficiently done by dynamic programming.

With rich local structural information, the structured discriminative model can play a complementary role to the tree transduction model that focus more on global syntactic structures.

4 Joint Decoding

From now on the remaining issue is jointly decoding the components. Either part factorizes over local structures: rules for the tree transduction model and ngrams for the language model or structured discriminative model. We may build a large dynamic programming table to utilize this kind of locality. Unfortunately this is computationally impractical. It is mathematically equivalent to perform exact dynamic programming decoding of Equation 1, which would consume asymptotically $O(SRL^{2(n-1)V})$ [1] time for building the chart (Cohn and Lapata, 2009). Cohn and Lapata (2009) proposed a beam search approximation along with cube-pruning heuristics to reduce the time complexity down to $O(SRBV)$ [2].

[1] S, R, L and V denote respectively for the number of source tree nodes, the number of rules, size of target lexicon and number of variables involved in each rule.

[2] B denotes the beam width.

In this work we utilize the efficiency of independent decoding from the two components respectively and then combine their solutions according to certain standards. This naturally results in a dual decomposition (Rush et al., 2010) solution.

Dual decomposition has been applied in several natural language processing tasks, including dependency parsing (Koo et al., 2010), machine translation (Chang and Collins, 2011; Rush and Collins, 2011) and information extraction (Reichart and Barzilay, 2012). However, the strength of this inference strategy has seldom been noticed in researches on language generation tasks.

We briefly describe the formulation here.

4.1 Description

We denote the pure tree transduction part and the pure ngram part as $g(\mathbf{y})$ and $f(\mathbf{z})$ respectively. Then joint decoding is equivalent to solving:

$$\max_{\mathbf{y}\in\mathcal{Y},\mathbf{z}\in\mathcal{Z}} \quad g(\mathbf{y}) + f(\mathbf{z}) \tag{3}$$

$$\text{s.t.} \quad \mathbf{z}_{kt} = \mathbf{y}_{kt},\ \forall k \in \{1,...,n\},\ \forall t \in \{0,1\},$$

where $\mathbf{y}$ denotes a derivation which yields a final compression $\{\mathbf{y}_1,...,\mathbf{y}_m\}$. This derivation comes from a pure tree transduction model. $\mathbf{z}$ denotes the compression composed of $\{\mathbf{z}_1,...,\mathbf{z}_m\}$ from an ngram compression model. Without loss of generality, we consider $\mathbf{y}_k$ and $\mathbf{z}_k$ as indicators that take value 1 if the k's token of original sentence has been preserved in the compression and 0 if it has been deleted. In the constraints of problem 3, $\mathbf{y}_{kt}$ or $\mathbf{z}_{kt}$ denote indicator variables that take value 1 if $\mathbf{y}_k$ or $\mathbf{z}_k = t$ and 0 otherwise.

Let $L(u,\mathbf{y},\mathbf{z})$ be the Lagrangian of (3). Then the dual objective naturally factorizes into two parts that can be evaluated independently:

$$\begin{aligned} L(u) &= \max_{\mathbf{y}\in\mathcal{Y},\mathbf{z}\in\mathcal{Z}} L(u,\mathbf{y},\mathbf{z}) \\ &= \max_{\mathbf{y}\in\mathcal{Y},\mathbf{z}\in\mathcal{Z}} g(\mathbf{y}) + f(\mathbf{z}) + \sum_{k,t} u_{kt}(\mathbf{z}_{kt} - \mathbf{y}_{kt}) \\ &= \max_{\mathbf{y}\in\mathcal{Y}}(g(\mathbf{y}) - \sum_{k,t} u_{kt}\mathbf{y}_{kt}) + \\ &\quad \max_{\mathbf{z}\in\mathcal{Z}}(f(\mathbf{z}) + \sum_{k,t} u_{kt}\mathbf{z}_{kt}) \end{aligned}$$

With this factorization, Algorithm 1 tries to solve the dual problem $\min_u L(u)$ by alternatively decoding each component.

This framework is feasible and plausible in that the two subproblems (line 3 and line 4 in Algorithm 1) can be easily solved with slight modifications on the values of the original dynamic programming chart. Joint decoding of a pure tree transduction model and a structured discriminative model is almost the same.

Algorithm 1 Dual Decomposition Joint Decoding

1: **Initialization**: $u_k^{(0)} = 0,\ \forall k \in \{1,...,n\}$
2: **for** $i = 1$ to MAX_ITER **do**
3: $\quad \mathbf{y}^{(i)} \leftarrow \text{argmax}_{\mathbf{y}\in\mathcal{Y}}(g(\mathbf{y}) - \sum_{k,t} u_{kt}^{(i-1)}\mathbf{y}_{kt})$
4: $\quad \mathbf{z}^{(i)} \leftarrow \text{argmax}_{\mathbf{z}\in\mathcal{Z}}(f(\mathbf{z}) + \sum_{k,t} u_{kt}^{(i-1)}\mathbf{z}_{kt})$
5: $\quad$ **if** $\mathbf{y}_{kt}^{(i)} = \mathbf{z}_{kt}^{(i)}\ \forall k\ \forall t$ **then**
6: $\quad\quad$ **return** $(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})$
7: $\quad$ **else**
8: $\quad\quad u_{kt}^{(i)} \leftarrow u_{kt}^{(i-1)} - \delta_i(\mathbf{z}_{kt}^{(i)} - \mathbf{y}_{kt}^{(i)})$
9: $\quad$ **end if**
10: **end for**

The asymptotic time complexity of Algorithm 1 is $O(k(SRV + L^{2(n-1)}))$, where k denotes the number of iterations. This is a significant reduction of $O(SRL^{2(n-1)V})$ by directly solving the original problem and is also comparable to $O(SRBV)$ of conducting beam search decoding.

We apply a similar heuristic with Rush and Collins (2012) to set the step size $\delta_i = \frac{1}{t+1}$, where $t < i$ is the number of past iterations that increase the dual value. This setting decreases the step size only when the dual value moves towards the wrong direction. We limit the maximum iteration number to 50 and return the best primal solution $\mathbf{y}^{(i)}$ among all previous iterations for cases that do not converge in reasonable time.

5 Experiments

5.1 Baselines

The pure tree transduction model and the discriminative model naturally become part of our baselines for comparison [3]. Besides comparing our methods against the tree-transduction model with ngram scores by beam search decoding, we also compare them against the available previous work from Galanis and Androutsopoulos (2010). This state-of-the-art work adopts a two-stage method to rerank results generated by a discriminative maximum entropy model.

5.2 Data Preparation

We evaluated our methods on two standard corpora [4], refer to as Written and Spoken respectively.

[3] The pure ngram language model should not be considered here as it requires additional length constraints and in general does not produce competitive results at all merely by itself.

[4] Available at http://jamesclarke.net/research/resources

We split the datasets according to Table 1.

Table 1: Dataset partition (number of sentences)

Corpus	Training	Development	Testing
Written	1,014	324	294
Spoken	931	83	254

All tree transduction models require parallel parse trees with aligned leaves. We parsed all sentences with the Stanford Parser [5] and aligned sentence pairs with minimum edit distance heuristic [6]. Syntactic features of the discriminative model were also taken from these parse trees.

For systems involving ngram scores, we trained a trigram language model on the Reuters Corpus (Volume 1) [7] with modified Kneser-Ney smoothing, using the widely used tool SRILM [8].

5.3 Model Training

The training process of a tree transduction model followed similarly to Cohn and Lapata (2007) using structured SVMs (Tsochantaridis et al., 2005). The structured discriminative models were trained according to McDonald (2006).

5.4 Evaluation Metrics

We assessed the compression results by the F1-score of grammatical relations (provided by a dependency parser) of generated compressions against the gold-standard compression (Clarke and Lapata, 2006). All systems were controlled to produce similar compression ratios (CR) for fair comparison. We also reported manual evaluation on a sampled subset of 30 sentences from each dataset. Three unpaid volunteers with self-reported fluency in English were asked to rate every candidate. Ratings are in the form of 1-5 scores for each compression.

6 Results

We report test set performance of the structured discriminative model, the pure tree transduction (T3), Galanis and Androutsopoulos (2010)'s method (G&A2010), tree transduction with language model scores by beam search and the proposed joint decoding solutions.

[5] http://nlp.stanford.edu/software/lex-parser.shtml

[6] Ties were broken by always aligning a token in compression to its last appearance in the original sentence. This may better preserve the alignments of full constituents.

[7] http://trec.nist.gov/data/reuters/reuters.html

[8] http://www-speech.sri.com/projects/srilm/

Table 2 shows the compression ratios and F-measure of grammatical relations in average for each dataset. Table 3 presents averaged human rating results for each dataset. We carried out pairwise t-test to examine the statistical significance of the differences [9]. In both datasets joint decoding with dual decomposition solution outperforms other systems, especially when structured models involved. We can also find certain improvements of joint modeling with dual decomposition on the original beam search decoding of Equation 1, under very close compression ratios.

Joint decoding of pure tree transduction and discriminative model gives better performance than the joint model of tree transduction and language model. From Table 3 we can see that integrating discriminative model will mostly improve the preservation of important information rather than grammaticality. This is reasonable under the fact that the language model is trained on large scale data and will often preserve local grammatical coherence, while the discriminative model is trained on small but more compression specific corpora.

Table 2: Results of automatic evaluation. (†: sig. diff. from T3+LM(DD); *: sig. diff. from T3+Discr.(DD) for $p < 0.01$)

Written	CR(%)	GR-F1(%)
Discriminative	70.3	$52.4^{\dagger *}$
G&A2010	71.6	60.2^{*}
Pure Tree-Transduction	72.6	$52.3^{\dagger *}$
T3+LM (Beam Search)	70.4	58.8^{*}
T3+LM (Dual Decomp.)	70.7	60.5
T3+Discr. (Dual Decomp.)	71.0	**62.3**
Gold-Standard	71.4	100.0

Spoken	CR(%)	GR-F1(%)
Discriminative	69.5	$50.6^{\dagger *}$
G&A2010	71.7	59.2^{*}
Pure Tree-Transduction	73.6	$53.8^{\dagger *}$
T3+LM (Beam Search)	75.5	59.5^{*}
T3+LM (Dual Decomp.)	75.3	61.5
T3+Discr. (Dual Decomp.)	74.9	**63.3**
Gold-Standard	72.4	100.0

Table 4 shows some examples of compressed sentences produced by all the systems in comparison. The two groups of outputs are compressions of one sentence from the Written corpora and the Spoken corpora respectively. Ungrammatical compressions can be found very often by several baselines for different reasons, such as the outputs from pure tree transduction and the discriminative model in the first group. The reason behind the

[9] For all multiple comparisons in this paper, significance level was adjusted by the Holm-Bonferroni method.

Table 3: Results of human rating. (†: sig. diff. from T3+LM(DD); *: sig. diff. from T3+Discr.(DD), for $p < 0.01$)

Written	**GR.**	**Imp.**	**CR(%)**
Discriminative	$3.92^{\dagger *}$	$3.46^{\dagger *}$	70.6
G&A2010	$4.11^{\dagger *}$	$3.50^{\dagger *}$	72.4
Pure Tree-Transduction	$3.85^{\dagger *}$	$3.42^{\dagger *}$	70.1
T3+LM (Beam Search)	$4.22^{\dagger *}$	3.69^{*}	73.0
T3+LM (Dual Decomp.)	**4.63**	3.98	73.2
T3+Discr. (Dual Decomp.)	4.62	**4.25**	73.5
Gold-Standard	4.89	4.76	72.9

Spoken	**GR.**	**Imp.**	**CR(%)**
Discriminative	$3.95^{\dagger *}$	$3.62^{\dagger *}$	71.2
G&A2010	$4.09^{\dagger *}$	3.96^{*}	72.5
Pure Tree-Transduction	$3.92^{\dagger *}$	$3.55^{\dagger *}$	71.4
T3+LM (Beam Search)	4.20^{*}	3.78^{*}	75.0
T3+LM (Dual Decomp.)	4.35	4.18	74.5
T3+Discr. (Dual Decomp.)	**4.47**	**4.26**	74.7
Gold-Standard	4.83	4.80	73.1

Table 4: Example outputs

Original: *It was very high for people who took their full-time education beyond the age of 18 , and higher among women than men for all art forms except jazz and art galleries .*
Discr.: *It was high for people took education higher among women .*
(Galanis and Androutsopoulos, 2010): *It was high for people who took their education beyond the age of 18 , and higher among women .*
Pure T3: *It was very high for people who took .*
T3+LM-BeamSearch: *It was very high for people who took their education beyond the age of 18 , and higher among women than men .*
T3+LM-DualDecomp: *It was very high for people who took their education beyond the age of 18 , and higher among women than men .*
T3+Discr.: *It was high for people who took education beyond the age of 18 , and higher among women than men .*
Gold-Standard: *It was very high for people who took full-time education beyond 18 , and higher among women for all except jazz and galleries .*

Original: *But they are still continuing to search the area to try and see if there were , in fact , any further shooting incidents .*
Discr.: *they are continuing to search the area to try and see if there were , further shooting incidents .*
(Galanis and Androutsopoulos, 2010): *But they are still continuing to search the area to try and see if there were , in fact , any further shooting incidents .*
Pure T3: *they are continuing to search the area to try and see if there were any further shooting incidents .*
T3+LM-BeamSearch: *But they are continuing to search the area to try and see if there were , in fact , any further shooting incidents .*
T3+LM-DualDecomp: *But they are continuing to search the area to try and see if there were any further shooting incidents .*
T3+Discr.: *they are continuing to search the area to try and see if there were further shooting incidents .*
Gold-Standard: *they are continuing to search the area to see if there were any further incidents .*

under generation of pure tree transduction is that it mainly deals with global syntactic integrity merely in terms of the application of synchronous rules. Introducing language model scores will smooth the candidate compressions and avoid many aggressive decisions of tree transduction. Discriminative models are good at local decisions with poor consideration of grammaticality. We can see that the joint models have collected their predictive power together. Unfortunately we can still observe some redundancy from our outputs in the examples. The size of training corpus is not large enough to provide enough lexicalized information.

On the other hand, the time consumption of the joint model with dual decomposition decoding in our experiments matched the aforementioned asymptotic analysis. The training process based on new decoding method consumes similar time as beam search with cube-pruning heuristic.

7 Conclusion and Future Work

In this paper we propose a joint decoding scheme for tree transduction based sentence compression. Experimental results suggest that the proposed framework works well. The overall performance gets further improved under our framework by introducing the structured discriminative model.

As several recent efforts have focused on extracting large-scale parallel corpus for sentence compression (Filippova and Altun, 2013), we would like to study how larger corpora can affect tree transduction and our joint decoding solution. Meanwhile, We would like to explore on how other text-rewriting problems can be formulated as a joint model and be applicable to similar strategies described in this work.

Acknowledgements

This work was supported by National Hi-Tech Research and Development Program (863 Program) of China (2014AA015102, 2012AA011101) and National Natural Science Foundation of China (61170166, 61331011). We also thank the anonymous reviewers for very helpful comments.

The contact author of this paper, according to the meaning given to this role by Peking University, is Xiaojun Wan.

References

Yin-Wen Chang and Michael Collins. 2011. Exact decoding of phrase-based translation models through lagrangian relaxation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 26–37, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. *computational linguistics*, 33(2):201–228.

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 the 44th annual meeting of the Association for Computational Linguistics*, pages 377–384. Association for Computational Linguistics.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:273–381.

Trevor Cohn and Mirella Lapata. 2007. Large margin synchronous generation and its application to sentence compression. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 73–82, Prague, Czech Republic, June. Association for Computational Linguistics.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 137–144. Association for Computational Linguistics.

Trevor Cohn and Mirella Lapata. 2009. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA, October. Association for Computational Linguistics.

Dimitrios Galanis and Ion Androutsopoulos. 2010. An extractive supervised two-stage method for sentence compression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 885–893, Los Angeles, California, June. Association for Computational Linguistics.

Minlie Huang, Xing Shi, Feng Jin, and Xiaoyan Zhu. 2012. Using first-order logic to compress sentences. In *AAAI*.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization-step one: Sentence compression. In *AAAI/IAAI*, pages 703–710.

Terry Koo, Alexander M. Rush, Michael Collins, Tommi Jaakkola, and David Sontag. 2010. Dual decomposition for parsing with non-projective head automata. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1288–1298, Cambridge, MA, October. Association for Computational Linguistics.

Ryan T McDonald. 2006. Discriminative sentence compression with soft syntactic evidence. In *EACL*.

Roi Reichart and Regina Barzilay. 2012. Multi-event extraction guided by global constraints. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 70–79, Montréal, Canada, June. Association for Computational Linguistics.

Alexander M. Rush and Michael Collins. 2011. Exact decoding of syntactic translation models through lagrangian relaxation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 72–82, Portland, Oregon, USA, June. Association for Computational Linguistics.

Alexander M Rush and Michael Collins. 2012. A tutorial on dual decomposition and lagrangian relaxation for inference in natural language processing. *Journal of Artificial Intelligence Research*, 45:305–362.

Alexander M Rush, David Sontag, Michael Collins, and Tommi Jaakkola. 2010. On dual decomposition and linear programming relaxations for natural language processing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1–11, Cambridge, MA, October. Association for Computational Linguistics.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. In *Journal of Machine Learning Research*, pages 1453–1484.

Katsumasa Yoshikawa, Tsutomu Hirao, Ryu Iida, and Manabu Okumura. 2012. Sentence compression with semantic role constraints. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 349–353. Association for Computational Linguistics.