Rule Selection with Soft Syntactic Features for String-to-Tree Statistical Machine Translation

Fabienne Braune and **Nina Seemann** and **Alexander Fraser**
CIS, Ludwig-Maximilians-Universität München
Oettingenstraße 67, 80538 München, Germany
[braunefe|seemanna]@ims.uni-stuttgart.de
fraser@cis.uni-muenchen.de

Abstract

In syntax-based machine translation, rule selection is the task of choosing the correct target side of a translation rule among rules with the same source side. We define a discriminative rule selection model for systems that have syntactic annotation on the target language side (string-to-tree). This is a new and clean way to integrate soft source syntactic constraints into string-to-tree systems as features of the rule selection model. We release our implementation as part of Moses.

1 Introduction

Syntax-based machine translation is well known for its ability to handle non-local reordering. Syntax-based models either use linguistic annotation on the source language side (Huang, 2006; Liu et al., 2006), target language side (Galley et al., 2004; Galley et al., 2006) or are syntactic in a structural sense only (Chiang, 2005). Recent shared tasks have shown that systems integrating information on the target language side, also called *string-to-tree* systems, achieve the best performance on several language pairs (Bojar et al., 2014). At the same time, soft syntactic features significantly improve the translation quality of hierarchical systems (Hiero) as shown in (Marton et al., 2012; Chiang, 2010; Liu et al., 2011; Cui et al., 2010). Improving the performance of string-to-tree systems through the integration of soft syntactic constraints on the source language side is therefore an interesting task.

So far, all approaches on this topic include soft syntactic constraints into the rules of string-to-tree (Zhang et al., 2011; Huck et al., 2014) or string-to-dependency (Huang et al., 2013) systems and define heuristics to determine to what extent these constituents match the syntactic structure of the source sentence. We propose a novel way to integrate soft syntactic constraints into a string-to-tree system. We define a discriminative rule selection model for string-to-tree machine translation. We consider rule selection as a multi-class classification problem where the task is to select the correct target side of a rule given its source side as well as contextual information about the source sentence and the considered rule. So far, such models have been applied to systems without syntactic annotation on the target language side. He et al. (2008), He et al. (2010) and Cui et al. (2010) apply such rule selection models to hierarchical machine translation, Liu et al. (2008) to tree-to-string systems and Zhai et al. (2013) to systems based on predicate argument structures. When target side syntactic annotations are taken into account, the task of rule selection has to be reformulated (see Section 2) while the same type of model can be used in approaches without target annotations. This work is the first attempt to define a rule selection model for a string-to-tree system. We make our implementation publicly available as part of Moses.[1]

We show in Section 2 that string-to-tree rule selection is different from the hierarchical case addressed by previous work and define our rule selection model. In Section 3 we present the training procedure before providing a proof-of-concept evaluation in Section 4.

2 Rule selection for string-to-tree SMT

2.1 String-to-tree machine translation

We present string-to-tree machine translation as implemented in Moses (which is the framework that we use). String-to-tree rules have the form $X/A \rightarrow \langle \alpha, \gamma, \sim \rangle$. On the source language side,

[1] We use the string-to-tree component of Moses (Williams and Koehn, 2012; Hoang et al., 2009) in which we integrate the high-speed classifier Vowpal Wabbit http://hunch.net/~vw/.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1095–1101,
Lisbon, Portugal, 17-21 September 2015.

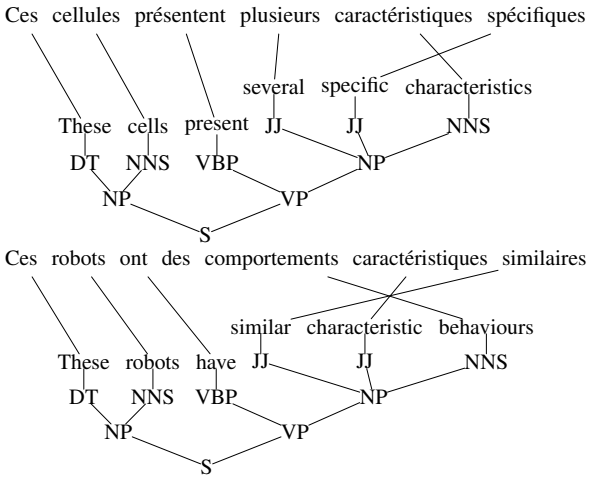

Figure 1: Word-aligned sentence pairs with target-side parse.

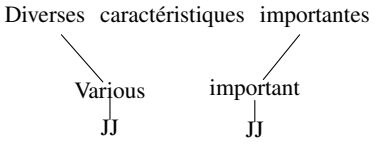

Figure 2: Partial translation during decoding.

all non-terminals have the unique label X while on the target language side non-terminals are annotated with syntactic labels $n_t \in N_t$. The left-hand side X/A consists of source and target non-terminals. In the right hand side (rhs), α is a string of source terminal symbols and the non-terminal X. The string γ consists of target terminals and non-terminals $n_t \in N_t$. The alignment $\sim$ is a one-to-one correspondence between source and target non-terminal symbols. String-to-tree rules are extracted from pairs of strings and trees as exemplified in Figure 1. Rules r_1 and r_2 are example rules extracted from this data.

(r_1) $X/NP \rightarrow \langle\ X_1$ caractéristiques X_2, $JJ_1\ JJ_2$ characteristics $\rangle$
(r_2) $X/NP \rightarrow \langle\ X_1$ caractéristiques X_2, NNS_1 characteristic $JJ_2\ \rangle$

During decoding, CYK+ chart parsing (Chappelier et al., 1998) with cube pruning and language model scoring is performed on an input sentence such as F below. Each time a rule is applied to the input sentence, candidate target trees are built. Figure 2 shows the partial translations built after the segments *Diverses* and *importantes* have been decoded. Given these partial translations, rule r_1 can be applied in a further decoding step.

F (Diverses)$_{X1}$ caractéristiques (importantes)$_{X2}$ n'ont pas été prises en compte.
(Various)$_{X_1}$ ***characteristics*** *(important)*$_{X_2}$ *were not considered.*

2.2 String-to-tree rule selection

Rule selection is the problem of selecting the rule with the correct target side among rules with the same source side. For hierarchical machine translation (Hiero), the rule selection problem consists of choosing, among r_3 and r_4, the rule that correctly applies to F (r_3 in our example).

(r_3) $X/X \rightarrow \langle\ X_1$ caractéristiques X_2, $X_1\ X_2$ characteristics $\rangle$
(r_4) $X/X \rightarrow \langle\ X_1$ caractéristiques X_2, X_1 characteristic $X_2\ \rangle$

Rule selection models disambiguate between these rules using context information about the source sentence and the shape of the rules.

In string-to-tree machine translation, the rule selection problem is different. Because the decoding process is guided by target side syntactic annotation, partial trees built during decoding must be considered when new rules are applied. For instance, when a rule is selected to translate sentence F given the partial translations in Figure 2, then the non-terminals in the target side of this rule must match the constituents selected so far. Consequently, rules r_1 and r_2 (Section 2.1) are not competing during rule selection.[2] Competing rules for r_1 would be r_5 and r_6 below.

(r_5) $X/NP \rightarrow \langle\ X_1$ caractéristiques X_2, JJ_1 properties $JJ_2\ \rangle$
(r_6) $X/NP \rightarrow \langle\ X_1$ caractéristiques X_2, $JJ_1\ JJ_2$ features $\rangle$

For consistency with decoding, we redefine the rule selection problem for the string-to-tree case. In this setup, it is the task of disambiguating rules with the same source side *and aligned target non-terminals*. As a consequence, our rule selection model (presented next) is not only normalized over the source rhs of the rules but also takes target non-terminals into account. The default rule scoring procedure for string-to-tree rules implemented in Moses uses the same normalization as we do. However, Williams and Koehn (2012) propose to normalize string-to-tree rules over the source rhs only.

[2]This is because their target side non-terminals are different.

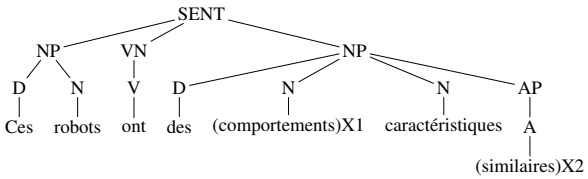

Figure 3: French sentence with input parse tree.

2.3 Rule selection model

We denote string-to-tree rules with $X/A \rightarrow \langle \alpha, \gamma, \sim \rangle$, as in Section 2.1. By $\tilde{N}t_t$, we denote target non-terminals with their alignment to source non-terminals.[3] $C(f, \alpha)$ is context information in the source sentence f and the source side α. $R(\alpha, \gamma)$ represents features on string-to-tree rules. The rule selection model estimates $P(\gamma \mid C(f, \alpha), R(\alpha, \gamma), \alpha, \tilde{N}t_t)$ and is normalized over the set G' of candidate target sides γ' for a given α *and* $\tilde{N}t_t$. The function $GTO : \alpha \rightarrow G'$ generates, given the source side α and target non-terminals $\tilde{N}t_t$, the set G' of all corresponding target sides γ'. The estimated distribution can be written as:

$$P(\gamma \mid C(f,\alpha), R(\alpha,\gamma), \alpha, \tilde{N}t_t) = \frac{\exp(\sum_i \lambda_i h_i(C(f,\alpha), R(\alpha,\gamma), \alpha, \tilde{N}t_t)))}{\sum_{\gamma' \in GTO(\alpha, \tilde{N}t_t)} \exp(\sum_i \lambda_i h_i(C(f,\alpha), R(\alpha,\gamma'), \alpha, \tilde{N}t_t))}$$

In the same fashion as (Cui et al., 2010) do for the hierarchical case, we define a global rule selection model instead of a model that is local to the source side of each rule.

To illustrate the feature templates $C(f, \alpha)$ and $R(\alpha, \gamma)$ of our rule selection model, we suppose that rule r_1 has been extracted from the French sentence in Figure 3. The syntactic features are:

- Does α match a constituent: *no_match*
- Type of matched constituent: *None*
- Lowest parent of unmatched constituent: *NP*
- Span width covered by α: *3*

The rule internal features are:

- Source side α: *X1_caractéristiques_X2* (one feature)
- Target side γ: *JJ1_JJ2_characteristics*
- Aligned terminals in α and γ: *caractéristiques↔characteristics*
- Aligned non-terminals in α and γ: *X1↔JJ1_X2↔JJ2*
- Best baseline translation probability: *Most_Frequent*

Our rule selection model is integrated in the Moses string-to-tree system as an additional feature of the log-linear model.

[3]For rule r_1,r_5 and r_6, $\tilde{N}t_t$ would be JJ_1 and JJ_2.

3 Model Training

We create training examples using the rule extraction procedure in (Williams and Koehn, 2012).[4] We begin by generating a rule-table using this procedure. Then, each time a rule $r : X/A \rightarrow \langle \alpha, \gamma, \sim \rangle$ can be extracted from the training data, we generate a new training example. The target side γ of the extracted rule is a positive instance and gets a loss of 0. To generate negative samples, we collect all rules $r_2, \ldots, r_n$ that have the same source language side as r as well as the same aligned target non-terminals $\tilde{N}t_t$. Each of these rules is a negative example and gets a cost of 1. As an example, suppose that rule r_1 introduced in Section 2.1 has been extracted from the training example in Figure 1. The target side "JJ_1 JJ_2 characteristics" is a correct class and gets a cost of 0. The target side of all other rules having the same source side *and aligned target non-terminals*, such as rule r_5 and r_6, are incorrect classes.

For model training, we use the cost-sensitive one-against-all-reduction (Beygelzimer et al., 2005) of Vowpal Wabbit (VW).[5] We avoid overfitting to training data by employing early stopping once classifier accuracy decreases on a held-out dataset.[6]

4 Experiments

4.1 Experimental Setup

Our baseline system is a syntax-based system with linguistic annotation on the target language side (*string-to-tree*). We use the version implemented in the Moses open source toolkit (Hoang et al., 2009; Williams and Koehn, 2012) with standard parameters. Rule extraction is performed as in (Galley et al., 2004) with rule composition (Galley et al., 2006; DeNeefe et al., 2007). Non-lexical unary rules are removed (Chung et al., 2011) and scope-3 pruning (Hopkins and Langmead, 2010) is performed. Rule scoring is done using relative frequencies normalized over the source rhs and aligned non-terminals in the target rhs. The contrastive system is the same string-to-tree system but augmented with our rule selection model as a feature of the log-linear model.

[4]Which is based on (Galley et al., 2004; Galley et al., 2006; DeNeefe et al., 2007).

[5]Specifically, the label dependent version of *Cost Sensitive One Against All* which uses classification.

[6]We use the development set which is also used for MIRA tuning.

System	science	medical	news
Baseline	34.06	49.87	18.35
Contrastive	34.36	49.57	18.59

Table 2: String-to-tree system evaluation results.

We evaluate the baseline and our global model on three domains: (1) news, (2) medical, and (3) science. The training data for news is taken from Europarl-v4. Development and test sets are from the news translation task of WMT 2009 (Callison-Burch et al., 2009). For medical we use the biomedical data from EMEA (Tiedemann, 2009). Since this is a parallel corpus only, we first removed duplicate sentences and then constructed development and test sets by randomly selecting sentence pairs. As training data for science we use the scientific abstracts data provided by Carpuat et al. (2013). Table 1 gives an overview of the corpora sizes.

Berkeley parser (Petrov et al., 2006) is used to parse the English side of each parallel corpus (for string-to-tree rule extraction) as well as for parsing the French source side (for feature extraction). We trained a 5-gram language model on the English side of each training corpus using the SRI Language Modeling Toolkit (Stolcke, 2002). We train the model in the standard way and generate word alignments using GIZA++. After training, we reduced the number of translation rules by only keeping the 30-best rules with the same source side according to the direct rule translation rule probability. Our rule selection model was trained with VW. All systems were tuned using batch MIRA (Cherry and Foster, 2012). We measured the overall translation quality with 4-gram BLEU (Papineni et al., 2002), which was computed on tokenized and lowercased data for all systems. Statistical significance is computed with the pairwise bootstrap resampling technique of Koehn (2004).

4.2 Results

Table 2 displays the BLEU scores for our experiments. On science and news, small improvements are achieved while for medical a small decrease is observed. None of these differences is statistically significant.

An analysis of the system outputs for each domain showed that the small improvements are due to the fact that in string-to-tree systems there is not enough ambiguity between competing rules during decoding. To support this conjecture, we first analyzed rule diversity by looking at the negative samples collected during training example acquisition. In a second step, we compared the results of the string-to-tree systems in Table 2 with a system where the translation rules are much more ambiguous. To this aim, we applied our approach to a hierarchical system in the same line as (Cui et al., 2010). Finally, we further tested the ability of our system to disambiguate between competing rules by training a model on the concatenation of all domains.

4.3 Analysis of Rule Diversity

The amount of competing rules during decoding can be estimated by looking at the negative samples collected for each training example. This analysis showed that the diversity of rules containing non-terminal symbols is limited. We present rules q_1 to q_3 (taken from science) to illustrate the poor diversity observed in our training examples.

(q_1) $X/PP \rightarrow \langle$ à X_1 X_2 éventail X_3, to DT_1 JJ_2 variety $PP_3 \rangle$
(q_2) $X/PP \rightarrow \langle$ à X_1 X_2 éventail X_3, to DT_1 JJ_2 range $PP_3 \rangle$
(q_3) $X/PP \rightarrow \langle$ à X_1 X_2 éventail X_3, to DT_1 JJ_2 array $PP_3 \rangle$

Rules q_1 to q_3 are the only rules with source side *à* X_1 X_2 *éventail* X_3. This number is very low given that the source side contains three non-terminal symbols out of which two are adjacent. Moreover, the difference between these rules is limited to the lexical translation of *éventail*. This lack of diversity is due to the constraint that competing string-to-tree rules must have the same aligned non-terminal symbols, which is taken into account when collecting negative samples. In other words, the ambiguity between translation rules in a string-to-tree system is heavily restricted by the target side syntax.

The observed lack of diversity could be minimized by allowing rules with the same source rhs to have different aligned target non-terminals. In this perspective, rule scoring should be done by normalizing over the source rhs only as in Williams and Koehn (2012). The rule selection model in Section 2.3 should then be redefined and normalized over all rules with the same source rhs. Another way to improve rule diversity would be to remove target non-terminals and use preference

	news	medical	science
training data	4th EuroParl corpus	(Tiedemann, 2009)	(Carpuat et al., 2013)
training data size	149,986 sentence pairs	111,081 sentence pairs	139,199 sentence pairs
development size	1,025 sentences	2,000 sentences	2,907 sentences
test size	1,026 sentences	1,999 sentences	3,915 sentences

Table 1: Overview of the sizes of the three domains.

System	science	medical	news
Baseline	31.22	48.67	17.28
Contrastive	**32.27**	**49.66**	17.38

Table 3: Hierarchical system evaluation results. The results in bold are statistically significant improvements over the Baseline (at confidence $p < 0.05$).

grammars as in Huck et al. (2014).

4.4 Comparison with Hierarchical Rule Selection

We applied our approach in a hierarchical phrase-based setting (Hiero). To this end, we trained 3 Hiero baseline systems and 3 Hiero systems augmented with our rule selection model on the data given in Section 4.1. The results of these experiments are shown in Table 3. Our augmented system largely outperforms the baselines. Interestingly, hierarchical rule selection significantly helps on the medical and scientific domain but still yields results that are significantly lower than those of the string-to-tree systems. This indicates that systems with target side syntax better disambiguate than hierarchical models with improved rule selection. Overall, we find the results of both types of systems promising and we will consider how to introduce more diversity into the rules of string-to-tree systems.

4.5 Concatenation of Training Data

In order to further evaluate the ability of our model to disambiguate string-to-tree rules, we trained a system using the concatenated training data of all 3 domains as presented in Section 4.1. This global model was then used to tune and decode using the development and test data of each domain. The results in Table 4 show that even on concatenated data our rule selection model does not improve over the baseline.

System	science	medical	news
Baseline	33.78	49.48	19.12
Contrastive	33.87	49.14	19.00

Table 4: String-to-tree system evaluation results with concatenated training data.

5 Conclusion and future work

We presented the first attempt to define a rule selection model with syntactic features for string-to-tree machine translation. We have shown that in order to be applied to the string-to-tree case, the rule selection problem must be redefined. An extensive evaluation on French-English translation tasks for different domains has shown that rule selection cannot significantly improve string-to-tree systems. An analysis of rule diversity and an empirical comparison with hierarchical rule selection indicate that the low improvements are due to the fact that the ambiguity between string-to-tree rules is too small to be improved with a rule selection model. In future work, we will use different techniques to improve the diversity of the string-to-tree rules considered during decoding in our system.

Acknowledgements

We thank all members of the DAMT team of the 2012 JHU Summer Workshop. We are especially grateful to Hal Daumé III and Ales Tamchyna for their ongoing support in the implementation of our system. We also thank Andreas Maletti for his shared expertise on tree grammars. This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 644402 (HimL) and the DFG grant *Models of Morphosyntax for Statistical Machine Translation (Phase 2)*, which we gratefully acknowledge.

References

Alina Beygelzimer, John Langford, and Bianca Zadrozny. 2005. Weighted one-against-all. In

AAAI, pages 720–725.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Ninth Workshop on Statistical Machine Translation*, WMT, pages 12–58, Baltimore, Maryland.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. 4th Workshop on Statistical Machine Translation*, pages 1–28.

Marine Carpuat, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013. Sensespotting: Never let your parallel data tie you to an old domain. In *Proc. ACL*.

Jean-Cédric Chappelier, Martin Rajman, et al. 1998. A generalized cyk algorithm for parsing stochastic cfg. *TAPD*, 98(133-137):5.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proc. NAACL*.

David Chiang. 2005. Hierarchical phrase-based translation. In *Proc. ACL*.

David Chiang. 2010. Learning to translate with source and target syntax. In *Proc. ACL*.

Tagyoung Chung, Licheng Fang, and Daniel Gildea. 2011. Issues concerning decoding with synchronous context-free grammars. In *Proc. ACL*.

Lei Cui, Dongdong Zhang, Mu Li, Ming Zhou, and Tiejun Zhao. 2010. A joint rule selection model for hierarchical phrase-based translation. In *Proc. ACL*.

Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. 2007. What can syntax-based mt learn from phrase-based mt. In *Proc. EMNLP*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. HLT-NAACL*.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve Deneefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. ACL*.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *Proc. COLING*.

Zhongjun He, Yao Meng, and Hao Yu. 2010. Maximum entropy based phrase reordering for hierarchical phrase-based translation. In *Proc. EMNLP*.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT*.

Mark Hopkins and Greg Langmead. 2010. Scfg decoding without binarization. In *Proc. EMNLP*.

Zhongqiang Huang, Jacob Devlin, and Rabih Zbib. 2013. Factored soft source syntactic constraints for hierarchical machine translation. In *Proc. EMNLP*.

Liang Huang. 2006. Statistical syntax-directed translation with extended domain of locality. In *In Proc. AMTA 2006*.

Mathias Huck, Hieu Hoang, and Philipp Koehn. 2014. Preference grammars and soft syntactic constraints for ghkm syntax-based statistical machine translation. In *Proc. SSST-8*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proc. ACL*.

Qun Liu, Zhongjun He, Yang Liu, and Shouxun Lin. 2008. Maximum entropy based rule selection model for syntax-based statistical machine translation. In *Proc. EMNLP*.

Lemao Liu, Tiejun Zhao, Chao Wang, and Hailong Cao. 2011. A unified and discriminative soft syntactic constraint model for hierarchical phrase-based translation. In *Proceedings of the 13th Machine Translation Summit*, pages 253–261.

Yuval Marton, David Chiang, and Philip Resnik. 2012. Soft syntactic constraints for arabic—english hierarchical phrase-based translation. *Machine Translation*, 26:137–157.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proc. ACL*.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken language Processing*.

Jörg Tiedemann. 2009. News from opus : A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing V*, volume V, pages 237–248. John Benjamins.

Philip Williams and Philipp Koehn. 2012. Ghkm rule extraction and scope-3 parsing in moses. In *Proc. WMT*.

Feifei Zhai, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2013. Handling ambiguities of bilingual predicate-argument structures for statistical machine translation. In *Proc. ACL*.

Jiajun Zhang, Feifei Zhai, and Chengqing Zong. 2011. Augmenting string-to-tree translation models with fuzzy use of source-side syntax. In *Proc. EMNLP*.