PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 105 APRIL 2016 101–110

Gibbs Sampling Segmentation of Parallel Dependency Trees for Tree-Based Machine Translation

David Mareček, Zdeněk Žabokrtský

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

We present a work in progress aimed at extracting translation pairs of source and target dependency treelets to be used in a dependency-based machine translation system. We introduce a novel unsupervised method for parallel tree segmentation based on Gibbs sampling. Using the data from a Czech-English parallel treebank, we show that the procedure converges to a dictionary containing reasonably sized treelets; in some cases, the segmentation seems to have interesting linguistic interpretations.

1. Introduction and related work

The context in which words and phrases are translated must be considered in machine translation. There are two basic ways how it is currently done in mainstream statistical machine translation (SMT). First, source-side sequences (phrases) longer than one word are stored together with their target-side equivalents in a "dictionary" (phrase table). Second, a language model rates possible longer sequences on the target side, which – among other things – reduces "boundary friction" between individually translated phrases. In addition, there are discriminative translation models that can profit from various types of features (including those from more distant context) too.

In dependency-tree-based MT, which constitutes the context of our study, the situation is more or less the same. Larger translation units (treelets composed of more than one node) can be used, like in Quirk et al. (2005). Target-side tree models (utilizing the probability of a word conditioned by its parent instead of its left neighbor(s)) can be used too to ensure that chosen target treelets fit together in the tree structure;

 Corresponding author: marecek@ufal.mff.cuni.cz
Cite as: David Mareček, Zdeněk Žabokrtský. Gibbs Sampling Segmentation of Parallel Dependency Trees for Tree-Based Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 105, 2016, pp. 101–110.
doi: 10.1515/pralin-2016-0005.

such a target-language dependency tree model was used in Žabokrtský and Popel (2009) (although the target tree model was combined only with a single-node translation model in this case). Third, the treelet translation model could be discriminative (i.e., capable of using more features from the context) too.

In this paper we focus on extracting a translation dictionary of pairs of source and target treelets from the node-aligned Czech-English parallel treebank CzEng.[1] We segment the trees into smaller parts called treelets. Then we produce a dictionary of (internally aligned) treelet pairs, equipped with source-to-target conditional probabilities (for both language directions) derived from treelet pair counts.[2]

Our approach is novel in two aspects:

- We use Gibbs sampling (Geman and Geman, 1984) for segmenting parallel trees, using a probabilistic model and a set of constraints that limit acceptable treelet pairs.
- We introduce interleaved trees, where nodes on odd levels contain lemmas of content words, whereas nodes on even levels[3] contain compact information on surface morphosyntactic form of the child node that is manifested in the surface sentence form.

The reasons why we use Gibbs sampling instead of exhaustive enumeration of all possible segmentations on both sides are the following. First, this approach leads to a relatively small translation dictionary, since it converges to segmentations that prefer repeated treelets (the rich-get-richer principle). Second, such a sampling approach allows us to describe only what the properties of the desired solutions are (in terms of a probabilistic model in combination with hard constraints on atomic sampling operations), and we do not need any specialized algorithms for finding such solutions – we just run the sampler. This seems to be a big advantage especially in the case of non-isomorphic trees and also because of noise caused by the fully automatic production of CzEng.

In the past, Bayesian methods (such as those based on Gibbs sampling or Pitman-Yor process) have been already used for tree segmentation. The typical purpose was grammar induction, both in constituency and dependency syntax, with Chung et al. (2014) being a representative of the former and Blunsom and Cohn (2010) of the latter. A dictionary of dependency treelet pairs, automatically extracted from parallel dependency trees, was used in the past too (e.g., Quirk et al., 2005; Ding and Palmer, 2004). However, to the best of our knowledge, there is no study merging these two worlds together. We are not aware of any attempt at finding a treelet translation dictionary for the needs of a real MT system using Gibbs sampling.

[1] All annotation contained in the treebank results from automatic tools like POS taggers, dependency parsers, and sentence and word aligners, see Bojar et al. (2012).

[2] Using the generated probabilistic treelet translation dictionary in a real MT system is left for further work. Interestingly, it seems that it will be possible to use Gibbs sampling also for decoding.

[3] The technical root added to each sentence is considered the first level.

Unlike the mainstream SMT, our approach relies on a fairly deep level of linguistic abstraction called tectogrammatical trees, as introduced by Sgall (1967), fully implemented for the first time in the Prague Dependency Treebank 2.0 (Hajič et al., 2006), and further adopted for the needs of tree-based MT in the TectoMT translation system (Žabokrtský et al., 2008). Only content words have nodes of their own in tectogrammatical trees, while function words disappear and are possibly turned to attributes inside the tectogrammatical nodes. Nodes of tectogrammatical trees are highly structured (they have tens of attributes, some of which further structured internally). Most of the attributes can be transferred from the source language to the target language relatively easily (for instance, the plural value of the grammatical number attribute goes most often to plural on the target side too). The attributes that are naturally most difficult to translate are *lemma* and *formeme* (the latter specification of the surface form, such as morphological case, or a function word such as a concrete preposition, or a verb clause type, see Dušek et al. (2012)). We follow Mareček et al. (2010) in using machine learning only for translating *lemmas* and *formemes*; the simpler-to-translate attributes are transferred by a less complex by-pass.

Since we want to keep the data structure used in the treelet transfer step as simple as possible, we convert tectogrammatical trees to so called *interleaved trees*, which contain only single-attribute nodes. Each original tectogrammatical node is split into a lemma node and a formeme node as the lemma's parent.[4] Regarding word-alignment, we only adopt the 1-to-1 alignment links from the original data.[5] In the interleaved trees, each such link is split into two: one connecting the *formeme* nodes and the other connecting the *lemma* nodes.

2. Segmentation by sampling

In order to generate a treelet translation dictionary, we need to split the aligned parallel trees from CzEng into smaller parts; we call them *bi-treelets*. Each bi-treelet consists of two subtrees (treelets) of the source and target trees respectively, and of alignment links internally connecting the two subtrees.

Virtually any tree edge can be cut across by the segmentation. However, since the source and the target trees are generally not isomorphic, we define additional constraints in order to receive technically reasonable bi-treelets.

- *Alignment constraint*: A pair of treelets has to be closed under alignment. In other words, no alignment link can refer outside of the bi-treelet.
- *Non-empty constraint*: Each bi-treelet must have at least one node both in the source and in the target tree. This constraint ensures that bi-trees projecting

[4] Valency of a governing word is usually determined by its lexeme (*lemma*), while the requirements imposed on its valency arguments are manifested by morphosyntactic features (*formemes*). Thus it seems more linguistically adequate to place the child's formeme between the parent and child's lemmas.

[5] We employed the links covered by the GIZA++ intersection symmetrization.

some nodes to nothing cannot exist and therefore both source and target dependency trees must be divided into the same number of treelets.

We use the Gibbs sampling algorithm to find the optimal translation bi-treelets. To model the probability of a segmented corpus, we use a generative model based on the Chinese restaurant process (Aldous, 1985). Assume that the corpus C is segmented to n bi-treelets $[B_1, \ldots, B_n]$. The probability that such a corpus is generated is

$$P(C) = p_t^{n-1}(1-p_t)\prod_{i=1}^{n}\frac{\alpha P_0(B_i) + count^{-i}(B_i)}{\alpha + i},$$

where $P_0(B_i)$ is a prior probability of a particular bi-treelet, hyperparameter α determines the strength of the prior, $count^{-i}(B_i)$ denotes how many times the bi-treelet B_i was generated before the position i, and p_t is the probability of generating the next bi-treelet.

The prior probability of a treelet is computed according to a separate generative micro-story: (1) We generate the node labels from a uniform distribution (probability $1/\#types$) and after each label, we decide whether to continue (probability p_c) or not ($1-p_c$), (2) When the labels are generated, we generate the shape of the tree from uniform distribution over all possible dependency trees with k nodes, which is k^{k-1}. This gives us the following formula for the treelet prior probability:

$$P_0(T) = \left(\frac{1}{\#types}\right)^k p_c^{k-1}(1-p_c)\frac{1}{k^{k-1}}$$

The bi-treelet prior probability is then a multiplication of the source and target treelet priors.[6]

Before sampling, we initialize bi-treelets randomly. We assign the binary attribute `is_segmented` to each dependency edge in both source and target trees. Technically, this attribute is assigned to the dependent node. Due to the alignment and non-empty constraints, the following conditions must be met:

- If two nodes are aligned, they must agree in the `is_segmented` attribute. In other words, both the nodes are roots of the bi-treelet or neither of them is.
- If two nodes are aligned, their closest aligned ancestors (parents, grandparents, etc.) should be aligned to each other. If not, there are some crossing alignment links, which could cause disconnected treelets during the sampling. To prevent this, the `is_segmented` attributes of such two nodes are permanently set to 1 and can not be changed during the sampling.
- If a node is not aligned, the `is_segmented` attribute is set permanently to 0 and cannot be changed during the sampling. This property connects all the not-aligned nodes to their closest aligned ancestors and ensure the non-empty constraint.

[6] We do not take into account possibly different alignment of nodes between the treelets.

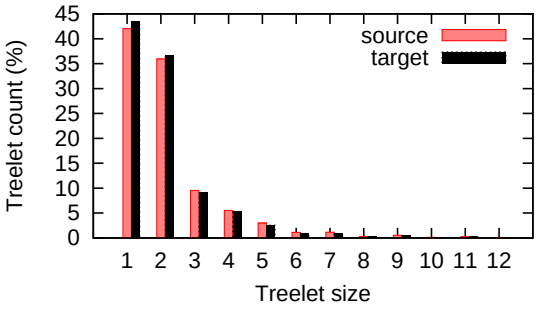

Figure 1. Distribution over different treelet sizes in the dictionary ($\alpha = 0.1$, $T = 1$, $c_p = 0.5$, $c_t = 0.99$).

The sampling algorithm goes through all the nodes in the source trees and samples a new binary value with respect to the corpus probability $P(C)$ (in case the change is not forbidden by the aforementioned constraints). The `is_segmented` attribute of its aligned counterpart in the target tree is set to the same value. Due to the exchangeability property, it is not necessary to compute the whole corpus probability. See the details in Cohn et al. (2009).

After a couple of "burn-in" iterations, the segmentation of trees converges to reasonable-looking bi-treelets. In the remaining iterations, the counts of bi-treelets are collected. Finally, the dictionary of bi-treelets with assigned probabilities computed from collected counts is created.

3. Experiments and evaluation

We perform our experiments on 10% of the Czech-English parallel treebank CzEng 1.0 (Bojar et al., 2012). This subset contains about 1.5 million sentences (21 million Czech tokens and 23 million English tokens) from different sources.

We started with initial setting of hyperparameters $\alpha = 0.1$, $p_c = 0.5$, and $p_t = 0.99$. The algorithm converges quite quickly. After the third iteration, the number of changes in the segmentation is less than 2% per iteration. Therefore we decided to set the "burn-in" period to the first 5 iterations and to start the collecting bi-treelets counts from the sixth iteration. The distribution over different sizes of treelets collected in the dictionary is depicted in Figure 1. There is more than 40% one-node treelets and about 35% two-node treelets. The average number of nodes in the bi-treelet is 2.07 in the source (English) and 1.99 in the target (Czech) side.

It is possible that for the decoding, we will need a dictionary with higher variance (more different treelets), so we use annealing to increase the number of segmentation

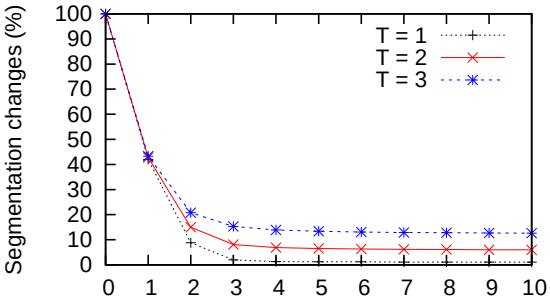

Figure 2. Percentage of changed segmentations during the first 10 iterations for different temperatures ($\alpha = 0.1$, $c_p = 0.5$, $c_t = 0.99$).

Continuation probability p_c	0.5	0.5	0.5	0.5	0.2	0.8	0.8
α	0.001	0.001	0.1	0.1	0.001	0.001	1
Temperature T	1	3	1	3	1	1	2
Last iteration dictionary size	2.45M	2.26M	2.48M	2.32M	2.49M	2.42M	2.34M
Collected dictionary size	2.69M	3.54M	2.73M	3.74M	2.58M	2.78M	3.31M
Average English treelet size	2.19	2.06	2.18	2.05	2.17	2.20	2.16
Average Czech treelet size	2.07	1.96	2.07	1.95	2.06	2.09	2.04

Table 1. The effect of setting the hyperparameters on the dictionary size and other quantities.

changes during the sampling. We introduce a temperature T and exponentiate all the probabilities by $1/T$. Temperatures higher than 1 flatten the distribution and boost the segmentation changes. Figure 2 shows that segmentation changes in the tenth iteration increased to 7% for $T = 2$ and to 12% for $T = 3$.

Table 1 shows the dictionary characteristics for different parameter settings. As expected, the collected dictionary size grows with growing temperatures, while the size of the dictionary based on the last iteration slightly decreases. Therefore, it will be easy to control the trade-off between the size of generated dictionary and the sharper distribution of translation candidates. Different values of the hyperparameter α do not affect the results much. Similarly, the continuation probability p_c does not affect the sizes of bi-treelets much.

We inspected the segmented trees after the last iteration; an example is shown in Figure 3. The thin edges are the ones cut by the segmentation, and the thick edges represent the delimited treelets (there are four bi-treelets in the figure). The lemma node and its respective formeme node often belong to the same treelet. Collocations

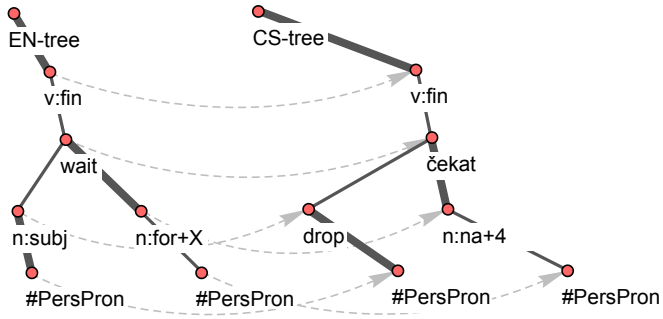

Figure 3. Interleaved trees representing the sentences “Čekal jsem na tebe.” – “I’ve been waiting for you.” and their segmentation to bi-treelets.

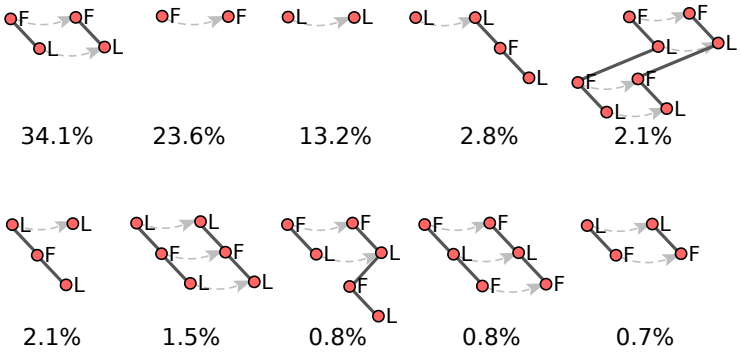

Figure 4. Distribution of the most frequent bi-treelets types in the dictionary (L = lemma, F = formeme).

(e.g. "European Union") also tend to constitute treelets of their own. The observation which we find most interesting is the manifestation of parallel verb valency captured by some treelets, such as the aligned formeme nodes *n:for+X – n:na+4* that are stuck to their governing verbs *wait – čekat* in a bi-treelet and not to their children.

Figure 4 shows 10 most frequent types of bi-treelets. We can see that if a pair of *formeme* nodes is inside a larger treelet it is connected to its respective pair of *lemma* nodes. Exceptions are the last two types of bi-treelets, where the *formeme* nodes are leaves. These are the cases of stronger valency between a parent *lemma* and morphosyntactic form of its dependent (e.g. *wait + n:for+X*).

4. Conclusions

We show a new method for obtaining a treelet-to-treelet translation dictionary from a parallel treebank using Gibbs sampling. In future work, we will evaluate our approach in a tree-based MT system.

Acknowledgments

The work was supported by the grant 14-06548P of the Czech Science Foundation. It has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

Bibliography

Aldous, David. Exchangeability and related topics. In *Ecole d'Ete de Probabilities de Saint-Flour XIII 1983*, pages 1–198. Springer, 1985.

Blunsom, Phil and Trevor Cohn. Unsupervised Induction of Tree Substitution Grammars for Dependency Parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1204–1213, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1870658.1870775`.

Bojar, Ondřej, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3921–3928, İstanbul, Turkey, 2012. European Language Resources Association. ISBN 978-2-9517408-7-7.

Chung, Tagyoung, Licheng Fang, Daniel Gildea, and Daniel Stefankovic. Sampling Tree Fragments from Forests. *Computational Linguistics*, 40(1):203–229, 2014. doi: 10.1162/COLI_a_00170. URL `http://dx.doi.org/10.1162/COLI_a_00170`.

Cohn, Trevor, Sharon Goldwater, and Phil Blunsom. Inducing compact but accurate tree-substitution grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 548–556, Boulder, Colorado, 2009.

Ding, Yuan and Martha Stone Palmer. Automatic Learning of Parallel Dependency Treelet Pairs. In Su, Keh-Yih, Jun ichi Tsujii, Jong-Hyeok Lee, and Oi Yee Kwong, editors, *IJCNLP*, volume 3248 of *Lecture Notes in Computer Science*, pages 233–243. Springer, 2004. ISBN 3-540-24475-1.

Dušek, Ondřej, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák, and David Mareček. Formemes in English-Czech Deep Syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 267–274, Montréal, Canada, 2012. Association for Computational Linguistics. ISBN 978-1-937284-20-6.

Geman, Stuart and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková-Razímová. Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 2006.

Mareček, David, Martin Popel, and Zdeněk Žabokrtský. Maximum Entropy Translation Model in Dependency-Based MT Framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–201, Uppsala, Sweden, 2010. Uppsala Universitet, Association for Computational Linguistics. ISBN 978-1-932432-71-8.

Quirk, Chris, Arul Menezes, and Colin Cherry. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 271–279, Stroudsburg, PA, USA, 2005. Association for

Computational Linguistics. doi: 10.3115/1219840.1219874. URL http://dx.doi.org/10.3115/1219840.1219874.

Sgall, Petr. *Generativní popis jazyka a česká deklinace*. Academia, Praha, 1967.

Žabokrtský, Zdeněk and Martin Popel. Hidden Markov Tree Model in Dependency-based Machine Translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 145–148, Suntec, Singapore, 2009. Association for Computational Linguistics. ISBN 978-1-932432-61-9.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, OH, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-09-1.

Address for correspondence:
David Mareček
marecek@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics,
Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1, Czech Republic