# Hidden Markov Tree Model in Dependency-based Machine Translation*

**Zdeněk Žabokrtský**
Charles University in Prague
Institute of Formal and Applied Linguistics
zabokrtsky@ufal.mff.cuni.cz

**Martin Popel**
Charles University in Prague
Institute of Formal and Applied Linguistics
popel@matfyz.cz

## Abstract

We would like to draw attention to Hidden Markov Tree Models (HMTM), which are to our knowledge still unexploited in the field of Computational Linguistics, in spite of highly successful Hidden Markov (Chain) Models. In dependency trees, the independence assumptions made by HMTM correspond to the intuition of linguistic dependency. Therefore we suggest to use HMTM and tree-modified Viterbi algorithm for tasks interpretable as labeling nodes of dependency trees. In particular, we show that the transfer phase in a Machine Translation system based on tectogrammatical dependency trees can be seen as a task suitable for HMTM. When using the HMTM approach for the English-Czech translation, we reach a moderate improvement over the baseline.

## 1 Introduction

Hidden Markov Tree Models (HMTM) were introduced in (Crouse et al., 1998) and used in applications such as image segmentation, signal classification, denoising, and image document categorization, see (Durand et al., 2004) for references.

Although Hidden Markov Models belong to the most successful techniques in Computational Linguistics (CL), the HMTM modification remains to the best of our knowledge unknown in the field.

The first novel claim made in this paper is that the independence assumptions made by Markov Tree Models can be useful for modeling syntactic trees. Especially, they fit dependency trees well, because these models assume conditional dependence (in the probabilistic sense) only along tree edges, which corresponds to intuition behind dependency relations (in the linguistic sense) in dependency trees. Moreover, analogously to applications of HMM on sequence labeling, HMTM can be used for labeling nodes of a dependency tree, interpreted as revealing the hidden states[1] in the tree nodes, given another (observable) labeling of the nodes of the same tree.

The second novel claim is that HMTMs are suitable for modeling the transfer phase in Machine Translation systems based on deep-syntactic dependency trees. Emission probabilities represent the translation model, whereas transition (edge) probabilities represent the target-language tree model. This decomposition can be seen as a tree-shaped analogy to the popular n-gram approaches to Statistical Machine Translation (e.g. (Koehn et al., 2003)), in which translation and language models are trainable separately too. Moreover, given the input dependency tree and HMTM parameters, there is a computationally efficient HMTM-modified Viterbi algorithm for finding the globally optimal target dependency tree.

It should be noted that when using HMTM, the source-language and target-language trees are required to be isomorphic. Obviously, this is an unrealistic assumption in real translation. However, we argue that tectogrammatical deep-syntactic dependency trees (as introduced in the Functional Generative Description framework, (Sgall, 1967)) are relatively close to this requirement, which makes the HMTM approach practically testable.

As for the related work, one can found a number of experiments with dependency-based MT in the literature, e.g., (Boguslavsky et al., 2004), (Menezes and Richardson, 2001), (Bojar, 2008). However, to our knowledge none of the published systems searches for the optimal target representa-

* The work on this project was supported by the grants MSM 0021620838, GAAV ČR 1ET101120503, and MŠMT ČR LC536. We thank Jan Hajič and three anonymous reviewers for many useful comments.

[1] HMTM looses the HMM's time and finite automaton interpretability, as the observations are not organized linearly. However, the terms "state" and "transition" are still used.



*Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 145–148,
Suntec, Singapore, 4 August 2009. 

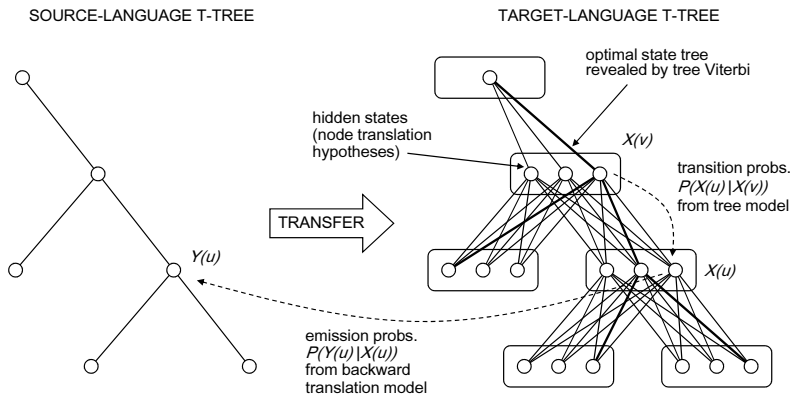


Figure 1: Tectogrammatical transfer as a task for HMTM.

tion in a way similar to HMTM.

## 2 Hidden Markov Tree Models

HMTM are described very briefly in this section. More detailed information can be found in (Durand et al., 2004) and in (Diligenti et al., 2003).

Suppose that $V = \{v_1, \ldots, v_{|V|}\}$ is the set of tree nodes, $r$ is the root node and $\rho$ is a function from $V \backslash r$ to $V$ storing the parent node of each non-root node. Suppose two sequences of random variables, $\boldsymbol{X} = (X(v_1), \ldots, X(v_{|V|}))$ and $\boldsymbol{Y} = (Y(v_1), \ldots, Y(v_{|V|}))$, which label all nodes from $V$. Let $X(v)$ be understood as a hidden state of the node $v$, taking a value from a finite state space $S = \{s_1, \ldots, s_K\}$. Let $Y(v)$ be understood as a symbol observable on the node $v$, taking a value from an alphabet $K = \{k_1, \ldots, k_2\}$. Analogously to (first-order) HMMs, (first-order) HMTMs make two independence assumptions: (1) given $X(\rho(v))$, $X(v)$ is conditionally independent of any other nodes, and (2) given $X(v)$, $Y(v)$ is conditionally independent of any other nodes. Given these independence assumptions, the following factorization formula holds:[2]

$$P(\boldsymbol{Y}, \boldsymbol{X}) = P(Y(r)|X(r))P(X(r)) \cdot \prod_{v \in V \backslash r} P(Y(v)|X(v))P(X(v)|X(\rho(v))) \quad (1)$$

We see that HMTM (analogously to HMM, again) is defined by the following parameters:

- $P(X(v)|X(\rho(v)))$ – transition probabilities between the hidden states of two tree-adjacent nodes,[3]
- $P(Y(v)|X(v))$ – emission probabilities.

Naturally the question appears how to restore the most probable hidden tree labeling given the observed tree labeling (and given the tree topology, of course). As shown in (Durand et al., 2004), a modification of the HMM Viterbi algorithm can be found for HMTM. Briefly, the algorithm starts at leaf nodes and continues upwards, storing in each node for each state and each its child the optimal downward pointer to the child's hidden state. When the root is reached, the optimal state tree is retrieved by downward recursion along the pointers from the optimal root state.

## 3 Tree Transfer as a Task for HMTM

**HMTM Assumptions from the MT Viewpoint.** We suggest to use HMTM in the conventional tree-based analysis-transfer-synthesis translation scheme: (1) First we analyze an input sentence to a certain level of abstraction on which the sentence representation is tree-shaped. (2) Then we use HMTM-modified Viterbi algorithm for creating the target-language tree from the source-language tree. Labels on the source-language nodes are treated as emitted (observable) symbols, while labels on the target-language nodes are understood as hidden states which are being searched for

[2] In this work we limit ourselves to *fully stationary* HMTMs. This means that the transition and emission probabilities are independent of $v$. This "node invariance" is an analogy to HMM's time invariance.

[3] The need for parametrizing also $P(X(r))$ (prior probabilites of hidden states in the root node) can be avoided by adding an artificial root whose state is fixed.

(Figure 1). (3) Finally, we synthesize the target-language sentence from the target-language tree.

In the HMTM transfer step, the HMTM emission probabilities can be interpreted as probabilities from the "backward" (source given target) node-to-node translation model. HMTM transition probabilities can be interpreted as probabilities from the target-language tree model. This is an important feature from the MT viewpoint, since the decomposition into *translation model* and *language model* proved to be extremely useful in statistical MT since (Brown et al., 1993). It allows to compensate the lack of parallel resources by the relative abundance of monolingual resources.

Another advantage of the HMTM approach is that it allows us to disregard the ordering of decisions made with the individual nodes (which would be otherwise nontrivial, as for a given node there might be constraints and preferences coming both from its parent and from its children). Like in HMM, it is the notion of hidden states that facilitates "summarizing" distributed information and finding the global optimum.

On the other hand, there are several limitations implied by HMTMs which we have to consider before applying it to MT: (1) There can be only one labeling function on the source-language nodes, and one labeling function on the target-language nodes. (2) The set of hidden states and the alphabet of emitted symbols must be finite. (3) The source-language tree and the target-language tree are required to be isomorphic. In other words, only node labeling can be changed in the transfer step.

The first two assumption are easy to fulfill, but the third assumption concerning the tree isomorphism is problematic. There is no known linguistic theory guaranteeing identically shaped tree representations of a sentence and its translation. However, we would like to show in the following that the tectogrammatical layer of language description is close enough to this ideal to make the HMTM approach practically applicable.

**Why Tectogrammatical Trees?** Tectogrammatical layer of language description was introduced within the Functional Generative Description framework, (Sgall, 1967) and has been further elaborated in the Prague Dependency Treebank project, (Hajič and others, 2006).

On the tectogrammatical layer, each sentence is represented as a tectogrammatical tree (t-tree for short; abbreviations t-node and t-layer are used in the further text too). The main features of t-trees (from the viewpoint of our experiments) are following. Each sentence is represented as a dependency tree, whose nodes correspond to autosemantic (meaningful) words and whose edges correspond to syntactic-semantic relations (dependencies). The nodes are labeled with the lemmas of the autosemantic words. Functional words (such as prepositions, auxiliary verbs, and subordinating conjunctions) do not have nodes of their own. Information conveyed by word inflection or functional words in the surface sentence shape is represented by specialized semantic attributes attached to t-nodes (such as number or tense).

T-trees are still language specific (e.g. because of lemmas), but they largely abstract from language-specific means of expressing non-lexical meanings (such as inflection, agglutination, functional words). Next reason for using t-trees as the transfer medium is that they allow for a natural transfer factorization. One can separate the transfer into three relatively independent channels:[4] (1) transfer of lexicalization (stored in t-node's lemma attribute), (2) transfer of syntactizations (stored in t-node's formeme attribute),[5] and (3) transfer of semantically indispensable grammatical categories[6] such as number with nouns and tense with verbs (stored in specialized t-node's attributes).

Another motivation for using t-trees is that we believe that local tree contexts in t-trees carry more information relevant for correct lexical choice, compared to linear contexts in the surface sentence shapes, mainly because of long-distance dependencies and coordination structures.

**Observed Symbols, Hidden States, and HMTM Parameters.** The most difficult part of the tectogrammatical transfer step lies in transfer-

[4]Full independence assumption about the three channels would be inadequate, but it can be at least used for smoothing the translation probabilities.

[5]Under the term syntactization (the second channel) we understand morphosyntactic form – how the given lemma is "shaped" on the surface. We use the t-node attribute *formeme* (which is not a genuine element of the semantically oriented t-layer, but rather only a technical means that facilitates modeling the transition between t-trees and surface sentence shapes) to capture syntactization of the given t-node, with values such as n:subj – semantic noun (s.n.) in subject position, n:for+X – s.n. with preposition *for*, n:poss – possessive form of s.n., v:because+fin – semantic verb as a subordinating finite clause introduced by *because*), adj:attr – semantic adjective in attributive position.

[6]Categories only imposed by grammatical constraints (e.g. grammatical number with verbs imposed by subject-verb agreement in Czech) are disregarded on the t-layer.

ring lexicalization and syntactization (attributes lemma and formeme), while the other attributes (node ordering, grammatical number, gender, tense, person, negation, degree of comparison etc.) can be transferred by much less complex methods. As there can be only one input labeling function, we treat the following ordered pair as the observed symbol: $Y(v) = (L^{src}(v), F^{src}(v))$ where $L^{src}(v)$ is the source-language lemma of the node $v$ and $F^{src}(v)$ is its source-language formeme. Analogously, hidden state of node $v$ is the ordered couple $X(v) = (L^{trg}(v), F^{trg}(v))$, where $L^{trg}(v)$ is the target-language lemma of the node $v$ and $F^{trg}(v)$ is its target-language formeme. Parameters of such HMTM are then following:

$P(X(v)|X(\rho(v))) = P(L^{trg}(v), F^{trg}(v)|L^{trg}(\rho(v)), F^{trg}(\rho(v)))$ – probability of a node labeling given its parent labeling; it can be estimated from a parsed target-language monolingual corpus, and

$P(Y(v)|X(v)) = P(L^{src}(v), F^{src}(v)|L^{trg}(v), F^{trg}(v))$ – backward translation probability; it can be estimated from a parsed and aligned parallel corpus.

To summarize: the task of tectogrammatical transfer can be formulated as revealing the values of node labeling functions $L^{trg}$ and $F^{trg}$ given the tree topology and given the values of node labeling functions $L^{src}$ and $F^{src}$. Given the HMTM parameters specified above, the task can be solved using HMTM-modified Viterbi algorithm by interpreting the first pair as the hidden state and the second pair as the observation.

## 4 Experiment

To check the real applicability of HMTM transfer, we performed the following preliminary MT experiment. First, we used the tectogrammar-based MT system described in (Žabokrtský et al., 2008) as a baseline.[7] Then we substituted its transfer phase by the HMTM variant, with parameters estimated from 800 million word Czech corpus and 60 million word parallel corpus. As shown in Table 1, the HMTM approach outperforms the baseline solution both in terms of BLEU and NIST metrics.

| System | BLEU | NIST |
|---|---|---|
| baseline system | 0.0898 | 4.5672 |
| HMTM modification | 0.1043 | 4.8445 |

Table 1: Evaluation of English-Czech translation.

## 5 Conclusion

HMTM is a new approach in the field of CL. In our opinion, it has a big potential for modeling syntactic trees. To show how it can be used, we applied HMTM in an experiment on English-Czech tree-based Machine Translation and reached an improvement over the solution without HMTM.

[7] For evaluation purposes we used 2700 sentences from the evaluation section of WMT 2009 Shared Translation Task. http://www.statmt.org/wmt09/

## References


Igor Boguslavsky, Leonid Iomdin, and Victor Sizov. 2004. Multilinguality in ETAP-3: Reuse of Lexical Resources. In *Proceedings of Workshop Multilingual Linguistic Resources, COLING*, pages 7–14.

Ondřej Bojar. 2008. *Exploiting Linguistic Data in Machine Translation*. Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Peter E. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*.

Matthew Crouse, Robert Nowak, and Richard Baraniuk. 1998. Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902.

Michelangelo Diligenti, Paolo Frasconi, and Marco Gori. 2003. Hidden tree Markov models for document image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:2003.

Jean-Baptiste Durand, Paulo Goncalvès, and Yann Guédon. 2004. Computational methods for hidden Markov tree models - An application to wavelet trees. *IEEE Transactions on Signal Processing*, 52(9):2551–2560.

Jan Hajič et al. 2006. Prague Dependency Treebank 2.0. Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase based translation. In *Proceedings of the HLT/NAACL*.

Arul Menezes and Stephen D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the workshop on Data-driven methods in machine translation*, volume 14, pages 1–8.

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the 3rd Workshop on SMT, ACL*.