Deep-syntax TectoMT for English-Spanish MT

Gorka Labaka, Oneka Jauregi, Arantza Díaz de Ilarraza,
Michael Ustaszewski, Nora Aranberri and Eneko Agirre
IXA Group
Faculty of Computer Science
University of the Basque Country
Spain
{gorka.labaka, ojauregi002, a.diazdeilarraza,
mustaszewski001, nora.aranberri, e.agirre}@ehu.eus

Abstract

Deep-syntax approaches to machine translation have emerged as an alternative to phrase-based statistical systems, which seem to lack the capacity to address essential linguistic phenomena for translation. As an alternative, TectoMT is an open source framework for transfer-based MT which works at the deep tectogrammatical level and combines linguistic knowledge and statistical techniques. This work describes the development of machine translation systems for English-Spanish in both directions, leveraging on the modules for the English-Czech TectoMT system. We show that it is feasible to develop basic systems with relatively low effort in 9 months. Our evaluation shows that despite not yet being able to beat a phrase-based statistical system, the TectoMT architecture offers flexible customization options, which considerably increase the BLEU scores.

1 Introduction

Phrase-based machine translation (MT) systems have difficulty in capturing linguistic phenomena, such as long-distance grammatical cohesion. Syntax-based approaches have appeared as an alternative that can overcome this barrier more easily. Shallow approaches, however, seem still too restrictive and methods of deep linguistic analysis have been put forward as a tool to capture all the important parts of the meaning of the text. Efforts to build translation models around deep syntactic structure often move the level of linguistic abstraction a step deeper into semantic roles and relations, which should entail a simpler transfer step because of the greater structural similarity between the deep structures of the source and target languages as compared to the surface realizations; better generalization of the language as it operates on lemmas of content words and grammatical constructions are abstracted with their meaning captured by language-independent attributes; and improved grammaticality of the output given the explicit representation of target-side sentence structure.

TectoMT (Žabokrtský et al., 2008; Popel and Žabokrtský, 2010) has emerged as a potential architecture to develop such an approach, together with other deep-transfer systems such as Matxin (Mayor et al., 2011) and the one proposed by Gasser (2012). In contrast to those systems, TectoMT combines linguistic knowledge and statistical techniques, particularly during transfer, and it aims at transfer on the so-called tectogrammatical layer (Hajičová, 2000), a layer of deep syntactic dependency trees.

In this paper we present a description of the work done to develop a TectoMT system for both directions of English-Spanish, based on the existing English-Czech TectoMT system. In Section 2 we give an overview of the TectoMT architecture and the key linguistic concepts it is based on; in Section 3 we describe the analysis, transfer and synthesis stages, and highlight the upgrades and modifications carried out to develop the new language pair; in Section 4 we show an initial evaluation of the new prototypes; and finally, in Section 5 we draw conclusions and comment on the planned future work.

Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015), pages 55–63,
Praha, Czech Republic, 3–4 September 2015.

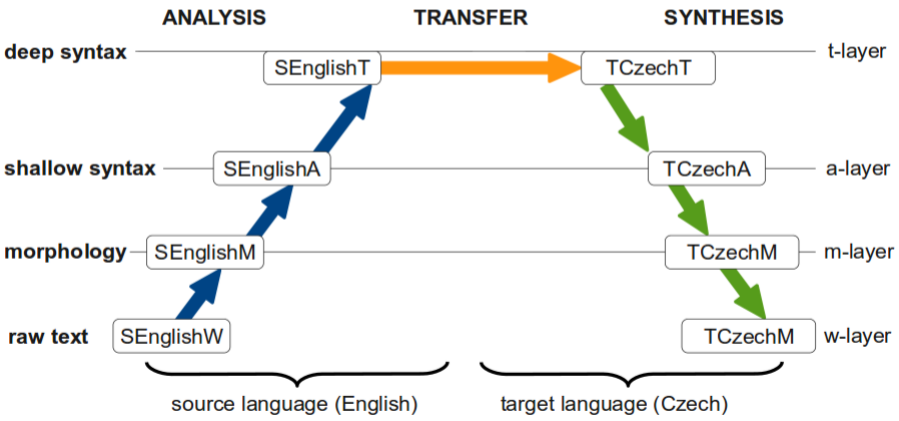

Figure 1: The general TectoMT architecture (from (Popel and Žabokrtský, 2010, :298)).

2 TectoMT architecture

As with most rule-based systems, TectoMT consists of an analysis, transfer and synthesis stages. The system works on different levels of abstraction (cf. Figure 1) and uses Blocks and Scenarios to process the information across the architecture.

2.1 Tecto layers

TectoMT works on an stratification approach to language, that is, it defines four layers of language, in increasing level of abstraction: raw text (word layer or w-layer), morphological layer (m-layer), shallow-syntax layer (analytical layer or a-layer), and deep-syntax layer (tectogrammatical layer or t-layer). This strategy is adopted from the Functional Generative Description theory (Sgall, 1967), which has been further elaborated and implemented in the Prague Dependency Treebank (PDT) (Hajič et al., 2006). As explained by (Popel and Žabokrtský, 2010, :296), each layer contains the following representation:

- **Morphological layer (m-layer)**

  Each sentence is tokenized and each token is annotated with a lemma and morphological tag.

- **Analytical layer (a-layer)**

  Each sentence is represented as a shallow-syntax dependency tree (a-tree). There is one-to-one correspondence between m-layer tokens and a-layer nodes. Each a-node is annotated with the type of dependency relation to its governing node or parent.

- **Tectogrammatical layer (t-layer)**

  Each sentence is represented as a deep-syntax dependency tree (t-tree). Autosemantic (meaningful) words are represented as t-layer nodes (t-nodes). Information conveyed by functional words (such as auxiliary verbs, prepositions and subordinating conjunctions) is represented by attributes of t-nodes. Most important attributes of t-nodes are:

  - tectogrammatical lemma;
  - functor: represents the semantic value of syntactic dependency relations, e.g. causal adjunct, conditional adjunct, actor, effect;
  - grammatemes: semantically oriented counterparts of morphological categories present at the higher level of abstraction, e.g. tense, number, verb modality, deontic modality, negation;
  - formemes: the morphosyntactic form of a t-node in the surface sentence. The set of formeme values compatible with a given t-node is limited by its semantic part of speech, e.g. subject noun, direct object noun, verb as a head of a relative clause (Dušek et al., 2012).

2.2 Blocks and Scenarios

Blocks are reusable components of subsequent steps into which NLP tasks can be decomposed. Each block has a well defined input and output specification and, in most cases, also a linguistically interpretable functionality. When developing new applications, blocks can be listed in a specific sequence and applied to the relevant data. These sequences are called scenarios.

TectoMT includes over a thousand blocks; approximately 224 blocks specific for English, 237 for Czech, over 57 for English-to-Czech transfer, 129 for other languages and 467 language-independent blocks.[1] Blocks vary in lengths, as they can consist of a few lines of code or tackle complex linguistic phenomena. To avoid code duplications, many routines are implemented separately and used in several blocks.

3 Development of a new language pair

We set to port the TectoMT system to work for the English-Spanish language pair in both directions. Because the original system covers both directions for the English-Czech pair, English analysis and synthesis were ready to use and our work mainly focused on Spanish analysis and synthesis, and on the transfer stages. In the following subsections we describe the work done on each step, analysis, transfer and synthesis, for each translation direction in our attempt to build tecto-level MT systems.

TectoMT is integrated within Treex,[2] a highly modular open source NLP framework implemented in Perl programming language. The framework includes modules for the English-Czech and Czech-English pairs, which are divided into language-specific and language independent blocks, thus facilitating the work to build the systems for the new language pair. As we will see in what follows, a good number of resources were reused, mainly those setting the general architecture and those specific to English; others were adapted, mainly those involving training of new language and translation models; and several new blocks were created to enable language-pair-specific features.

3.1 Analysis

The analysis stage aims at getting raw input text and analyzing it up to the tectogrammatical level so that transfer can be performed (cf. figs. 2 and 3). For English, the modules needed for analysis were already developed and running, and therefore little effort had to be put on it.

For Spanish, however, new analysis tools had to be integrated into Treex. For tokenization and sentence splitting, we adapted the modules of Treex to Spanish. Treex integrates tokenization and sentence splitting based on non-breaking prefixes. Therefore, we added a list of Spanish non-breaking prefixes in the module.

For the remaining tasks, we opted for the `ixa-pipes tools`.[3] These tools consist of a set of modules that perform linguistic analysis from tokenization to parsing, as well as several external tools that have been adapted to interact with them, adding extra functionality. We integrated the lemmatization and POS tagging (`ixa-pipe-pos`) and the dependency parsing (`ixa-pipe-srl`) tools in Treex. The first provides Perceptron (Collins, 2002) and Maximum Entropy (Ratnaparkhi, 1999) POS tagging models trained and evaluated using the AnCora corpus via 10-fold cross-validation, dictionary-based lemmatization, multiword detection and post-processing of probabilistic model pos tags using monosemic dictionaries. The second provides constituent parsing trained on the AnCora corpus and HeadFinders based on Collins head rules (Collins, 1999).

The tools were already developed, with accurate models for Spanish, and ready to use. Our efforts focused on their integration within Treex. We did this by adding them as wrapper blocks that, given a set of already tokenized sentences, creates the appropriate input in the corresponding format and calls the relevant tool. Once the tools complete their work, the output of the system is read and loaded in Treex documents.

[1] Statistics taken from: `https://github.com/ufal/treex.git(27/08/2015)`

[2] `https://ufal.mff.cuni.cz/treex`, `https://github.com/ufal/treex`

[3] `http://ixa2.si.ehu.es/ixa-pipes/`

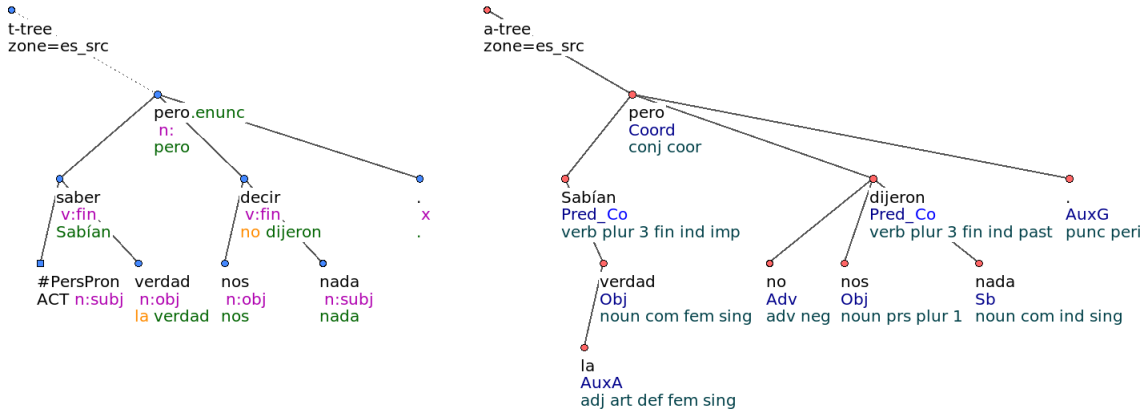

Figure 2: a-level and t-level Spanish analysis.

All `ixa-pipes tools` read NAF documents (with word forms and term elements) via standard input and output NAF through standard output. The NAF format is a linguistic annotation format designed for complex NLP pipelines (Fokkens et al., 2014).

The analyses generated by the `ixa-pipes tools` follow the AnCora guidelines both for morphological tags and dependency tree structures. This mostly equates to the a-layer in the TectoMT stratification. Therefore, to fully integrate the analyses into Treex and generate the expected a-tree, the analyses had to be mapped to a universal PoS and dependency tags. TectoMT currently uses the Interset tagset (Zeman, 2008) and HamleDT guidelines (Zeman et al., 2014). To implement this mapping, we used existing modules such as the Interset driver for Spanish AnCora Treebank tagset[4] by Dan Zeman and Zdenek Zabokrtsky, and the Harmonization Treex block for Spanish AnCora-style dependencies[5] by Dan Zeman, Zdenek Zabokrtsky and Martin Popel. On top of these, and in order to form the t-level tree, we used 16 additional blocks:

1. **Language-independent blocks.** 11 of the blocks were simply reused from the language-independent set already available in Treex. These mainly re-arrange nodes, mark heads (coordinations, clauses, coreference) and set node types.

2. **Adapted blocks.** 4 blocks were adapted from blocks originally used for English or Czech analysis. These include how to mark edges to collapse nodes into a single t-level node, how to annotate a number of functions words, sentence mood and grammateme values.

3. **New language-specific blocks.** 1 block was specifically written to set the grammatemes based on the Interset tagset features (and formemes) of the corresponding auxiliary a-level nodes.

3.2 Transfer

The transfer stage uses a statistical transfer dictionary together with a set of manually written blocks. The transfer dictionary is trained on parallel corpora analyzed up to the t-level in both languages. Learning equivalences at this level of representation enriches the model and simplifies the complexity of translation: it is not word-form equivalences that are learned, but rather the final dictionary includes the translation of lemmas, formemes and grammatemes (Žabokrtský, 2010). This approach is based on the assumption that t-tree structures in different languages are shared. Although this is not always true (Popel, 2009), it allows to model the working language pair as source-target one-to-one mapping.

For each t-lemma and formeme in a source t-tree, the translation model (TM) assigns a score to all possible translations observed in the training data. This score is a probability estimate of the translation

[4] `https://metacpan.org/source/ZEMAN/Lingua-Interset-2.041/lib/Lingua/Interset/Tagset/ES/Conll2009.pm`

[5] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/HamleDT/ES/Harmonize.pm`

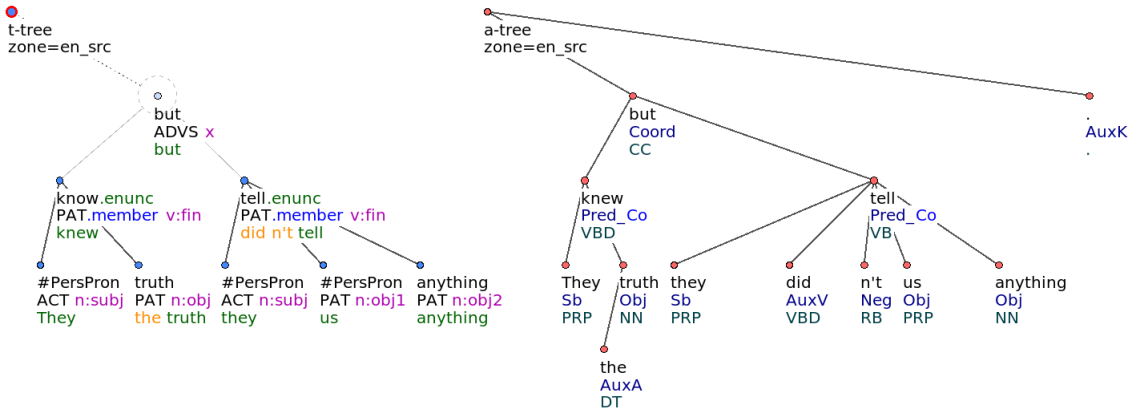

Figure 3: a-level and t-level English analysis.

variant given the source t-lemma and formeme, and other contextual information, and it is calculated as a linear combination of two main components:

- **The discriminative TM** (Mareček et al., 2010) is a set of maximum entropy (MaxEnt) models (Berger et al., 1996) trained for each specific source t-lemma and formeme, where the prediction is based on features extracted from the source tree (Crouse et al., 1998; Žabokrtský and Popel, 2009).

- **The dictionary TM** is a bilingual dictionary that contains a list of possible translation equivalents based on relative frequencies and no contextual features.

Both components are trained on the parallel corpora at the t-level. The final score assigned to each t-lemma and formeme in the TMs is calculated through interpolation. For the t-lemmas, weights of 0.5 and 1 are assigned to the dictionary TM and the discriminative TM, respectively. In the case of formemes, the values are reversed. Using these two TMs, we obtain a weighted n-best list of translation variants for each t-lemma and each formeme. The lists are jointly re-ranked by Hidden Markov Tree Models (HMTM), similarly to standard chains but operating on trees (Crouse et al., 1998; Žabokrtský and Popel, 2009). This setting was taken as-is from the one used for English-Czech.

The hybrid architecture of TectoMT, where both statistical transfer models and manually defined blocks can be combined, allows the integration of domain specific human dictionaries. Our development targets a question-and-answer (Q&A) scenario in the information technology (IT) domain. Therefore, in order to customize the systems to this domain, we integrated the Microsoft Terminology Collection as preprocessing (so the two TMs serve as a backoff for this human in-domain dictionary). The Microsoft Terminology Collection is freely available[6] and contains 22,475 entries.

The equivalence of grammatemes is assigned by manually written rules. The information they contain is linguistically more abstract, e.g. tense and number, and it is usually paralleled in the target language. Therefore, a set of relatively simple rules (with a list of exceptions) is sufficient for this task. These rules are inherently language-specific. At the time of writing, we use 5 blocks specifically written for the English-to-Spanish direction. These blocks address the lack of gender in English nouns (necessary in Spanish), differences in definiteness and articles, differences in structures such as *There is...* and relative clauses.

3.3 Synthesis

The output from transfer is a t-level tree that must be interpreted during the synthesis stage to generate the a-tree, which is used to create the final raw text (cf. figs. 4 and 5). The English synthesis was already developed and therefore, once again, our work mainly focused on preparing the Spanish synthesis, as we explain below.

[6] http://www.microsoft.com/Language/en-US/Terminology.aspx

We distinguish three steps during synthesis. On a first step, the t-tree generated using the information obtained during transfer must be transformed into an a-tree. At the time of writing, we use a total of 24 blocks.

1. **Language-independent blocks.** 9 of the blocks were reused from the language-independent set already available in Treex. Among these are blocks to mark subjects, impose subject-predicate and attribute agreements, add separate negation nodes, add specific punctuation for coordinate clauses, or impose capitalization at the beginning of sentence.

2. **Adapted blocks.** 12 blocks were adapted from the blocks in the English and Czech synthesis, or generic ones. For example, after acquiring the tree structure, the morphological categories are filled with values derived from the grammatemes and formemes. Whereas this is done for all languages, Spanish requires information coming from English grammatemes to be further distinguished. This is the case of the imperfect tense (a subcategory of past tense) and imperfect aspect, for instance, which we set on a block. Another block deals with articles. Knowing the definiteness of a noun or noun phrase is not always enough to decide whether to generate a determiner in the target language, and when necessary, to generate the appropriate one. Similarly, we adapted blocks for prepositions, subordinate conjunctions and auxiliary verbs. To mention yet another block, we remove personal pronoun nodes when acting as subject (the information is passed on to the predicate) as pro-drop languages such as Spanish do not require that they appear explicitly because this is already marked in the verb.

3. **New language-specific blocks.** 3 blocks were written from scratch to deal with Spanish-specific features. These deal with attribute order, comparatives and verb tenses. Attribute order refers to the position of adjectives with respect to the unit they modify. In English, adjectives occur before the noun they modify, but this is the opposite - with some exceptions for figurative effect - in Spanish. The block addressing comparatives creates additional nodes for the Spanish structure, which is specially relevant for the cases where no separate comparative word is used in English. Finally, a block was specifically written to address the complex verb tenses in Spanish. This block uses the information about tense, perfectiveness and progressiveness of the English verb to select the appropriate verb form in Spanish.

Overall, we see that most blocks are used (i) to fill in morphological attributes that will be needed in the second step, (ii) to add function words where necessary, (iii) to remove superfluous nodes, and (iv) to add punctuation nodes.

On a second step, the lemma and morphosyntactic information on the a-tree must be turned into word forms to generate the w-tree. We used Flect (Dušek and Jurčíček, 2013) to do this, by training new models for Spanish. Flect is a statistical morphological generation tool based on Python and Scikit-Learn that learns morphological inflection patterns from corpora. We trained the system with a subset of morphologically annotated Europarl corpus (530K tokens) where the system automatically learns how to generate inflected word forms from lemmas and morphological features. Flect can inflect previously unseen words as it uses lemma suffixes as features and predicts edit scripts that describe the difference between the lemma and the form, which improves robustness.

On a third step, once we obtain the w-tree with the word forms, a number of blocks can be written to polish the final output. For example, we use a block to concatenate the prepositions *a* and *de* with the masculine singular article *el*, which should be presented as the single forms *a*+*el* $\rightarrow$ *al* and *de*+*el* $\rightarrow$ *del*.

4 Evaluation

We evaluated the new English-to-Spanish and Spanish-to-English TectoMT prototypes in three different scenarios: using language-independent blocks only,[7], adding the blocks written and adapted for Spanish, and adding the domain-specific dictionary.

[7]This setup includes `ixa-pipes tools` and Flect models for Spanish analysis and synthesis, and bilingual transfer models.

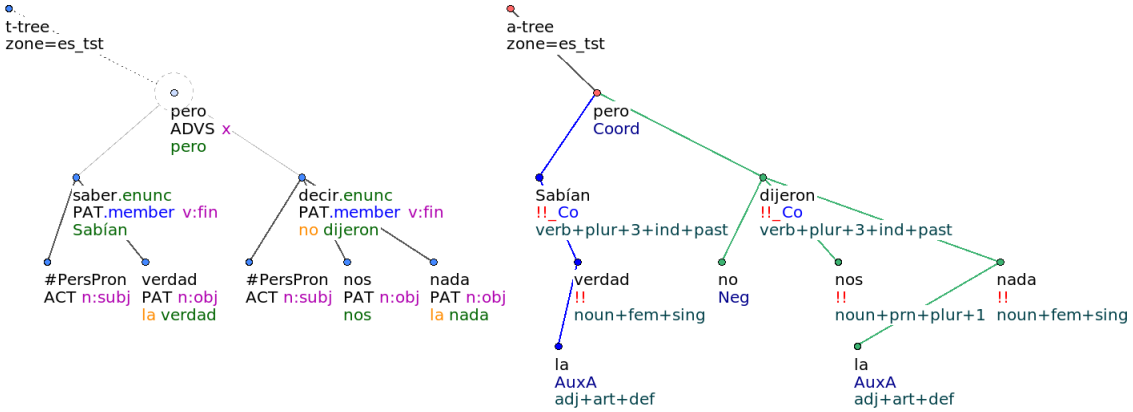

Figure 4: a-level and t-level Spanish synthesis.

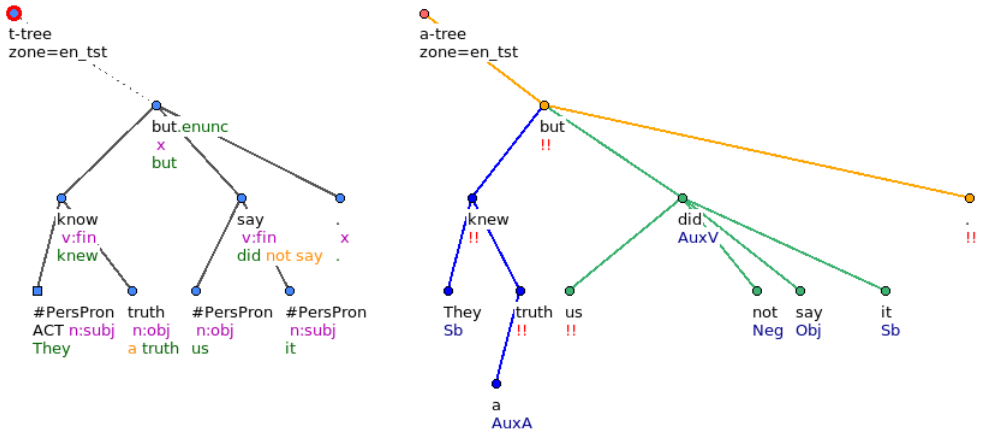

Figure 5: a-level and t-level English synthesis.

Also, we evaluated the new systems against a phrased-based statistical system. To this end, we built two SMT systems, one per language direction. We used tools available in the Moses toolkit for tokenization and truecasing, while mGiza was used for word alignment. For language modeling, we use SRILM to train a different language model (LM) for each corpus available and to combine them through LM interpolation. We trained the systems on bilingual corpora including Europarl, United Nations, News Commentary and Common Crawl (∼355 million words). The monolingual corpora used to learn the LM include the target-side texts of Europarl, News Commentary and News Crawl (∼60 million words). As previously mentioned, our efforts focus on a question-and-answer (Q&A) scenario in the information technology (IT) domain. Therefore, for tuning, we used a set of 1,000 in-domain interactions (question-answer pairs). The original interactions were in English and they were translated into Spanish by human translators.

We calculated BLEU scores for the systems on two different test sets. The first is another test-set of 1,000 in-domain interactions. The second is the newswire test-set used in the WMT11 campaign (Table 1) .

We can draw several conclusions from the BLEU scores obtained for each system and language pair. First, we observe that the TectoMT prototypes do not yet beat the statistical systems, although the English-to-Spanish system evaluated on the IT test-set remains very close, less than 2 BLEU points behind. Because a large portion of the TectoMT systems is based on rules, the lower scores of the baselines was to be expected given the effort put at this stage of development.

With regard to the TectoMT systems, we observe how the BLEU scores increase as we customize the system. The baseline systems with only language-independent blocks score lower than the systems

	English-Spanish		Spanish-English	
	IT	WMT11	IT	WMT11
Moses	28.12	26.91	31.92	25.24
TectoMT – language independent blocks	12.40	8.38	12.34	8.17
TectoMT – + Spanish blocks	23.62	13.92	14.67	8.50
TectoMT – + domain dictionary	26.40	13.25	15.82	8.23

Table 1: BLEU scores for the English-Spanish TectoMT prototypes

which include Spanish-specific blocks. For the English-to-Spanish system, BLEU scores almost double. For the Spanish-to-English system scores also increase although not as much. When adding the Microsoft dictionary (IT domain-specific), we observe that the BLEU scores increase almost 3 points for the English-to-Spanish direction and over 1 point for the Spanish-to-English direction. It is worth noting the small setback introduced by this specialized dictionary for the news domain with a drop of 0.67 and 0.27.

The scores also show the difference in development effort for the TectoMT systems in terms of language direction. The baseline TectoMT systems score similarly for both directions, at around 12 BLEU points for the IT test-set and 8 BLEU points for the WMT11 test-set. However, the priority given to Spanish-specific blocks for synthesis result in a better system for the English-to-Spanish direction.

Finally, it is worth mentioning the difference in scores between the test-sets, as the IT test-set scores substantially higher than the newswire test-set. This is probably because the IT domain test-set contains shorter and less convoluted sentences and most development work was based on IT-domain text analysis, even if the blocks written deal with generic linguistic features.

As a reference of the human effort required, we developed the new TectoMT systems over a period of 9 months.

5 Conclusions

In this paper we have shown the work done to develop entry-level deep-syntax systems for the English-Spanish language pair following the tectogrammatical MT approach. Thanks to previous work done for the English-Czech pair, we have reused most of the English analysis and synthesis modules, and mainly focused on the integration of tools and the development of models and blocks for Spanish. In particular, we have integrated the `ixa-pipes tools` for PoS and dependency parsing of Spanish, and adapted its output to comply with the tecto-level representation of language, which uses universal labels. For transfer, we have trained new statistical models for both English-to-Spanish and Spanish-to-English directions. For synthesis, we have trained a new morphological model to obtain Spanish word forms. Substantial effort was also put on writing sets of blocks to address differing linguistic features between the language pairs across all stages with a total of 55 reused blocks and 5 new/adapted blocks for the Spanish-to-English direction, and a total of 73 reused blocks and 19 new/adapted blocks for the English-to-Spanish direction. The system is open source and can be downloaded from `https://github.com/ufal/treex`. The evaluation has shown that the English-Spanish TectoMT prototype systems do not yet score as high as a phrase-based statistical system. However, the TectoMT architecture offers flexible customization options. We have shown that the BLEU scores can increase considerably as these are integrated and tuned to the working language pair.

Acknowledgements

The research leading to these results has received funding from FP7-ICT-2013-10-610516 (QTLeap).

References

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Michael Collins. 1999. Head-driven statistical models for natural language parsing.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

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

Ondrej Dušek and Filip Jurčíček. 2013. Robust multilingual statistical morphological generation models. *ACL 2013*, page 158.

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák, and David Mareček. 2012. Formemes in English-Czech deep syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 267–274. Association for Computational Linguistics.

Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert van Hage, and Piek Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 9, Reykjavik, Iceland.

Michael Gasser. 2012. Toward a rule-based system for English-Amharic translation. *Language Technology for Normalisation of Less-Resourced Languages*, page 41.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štepánek, Jiří Havelka, Marie Mikulová, Zdenek Žabokrtský, and Magda Ševcıková Razımová. 2006. Prague dependency treebank 2.0. *CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia*, 98.

Eva Hajičová. 2000. Dependency-based underlying-structure tagging of a very large Czech corpus. *TAL. Traitement automatique des langues*, 41(1):57–78.

David Mareček, Martin Popel, and Zdeněk Žabokrtský. 2010. Maximum entropy translation model in dependency-based MT framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–206. Association for Computational Linguistics.

Aingeru Mayor, Iñaki Alegria, Arantza Díaz De Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for Basque. *Machine translation*, 25(1):53–82.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: modular NLP framework. In *Advances in natural language processing*, pages 293–304. Springer.

Martin Popel. 2009. Ways to improve the quality of English-Czech machine translation. *Master's thesis, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic*.

Adwait Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. *Machine learning*, 34(1-3):151–175.

Petr Sgall. 1967. Functional sentence perspective in a generative description. *Prague studies in mathematical linguistics*, 2(203-225).

Zdeněk Žabokrtský and Martin Popel. 2009. Hidden markov tree model in dependency-based machine translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 145–148. Association for Computational Linguistics.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170. Association for Computational Linguistics.

Zdeněk Žabokrtský. 2010. From treebanking to machine translation. *Habilitation thesis, Charles University, Prague, Czech Republic*.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.

D. Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of LREC*, pages 213–218.