Chapter 36

Bootstrapping a dependency parser for Maltese – a real-world test case

Jörg Tiedemann
University of Helsinki

Lonneke van der Plas
University of Malta

This paper evaluates the practicality of methods intended to bootstrap dependency parsers for new languages on a real-world test case: Maltese. Previous work has evaluated cross-lingual methods, such as annotation projection and model transfer, by proxy, i.e., by selecting target languages from the set of languages available in multilingual treebanks, because truly under-resourced languages do not have test sets available. As a result, experiments in previous work are often limited to closely related Indo-European languages, or lack real-world scenarios. At this exact point in time, Maltese is an excellent candidate to evaluate the usefulness of the cross-lingual methods proposed in previous work: treebank development is in progress, no syntactic parsers are available, but certain NLP tools and corpora have recently become available. Maltese belongs to the branch of Semitic languages that is different from the languages for which NLP resources are most widely available. However, it has been under the influence of several indo-European language due to its turbulent history. It is therefore an even more interesting test case for exploring multi-source projection and the contribution of various languages with respect to their linguistic influence on the Maltese language.

1 Introduction

State-of-the-art methods for inducing NLP tools, such as part-of-speech (PoS) taggers and dependency parsers, rely on large quantities of hand-annotated data. For most of the world's languages these annotations are not available, because their creation is a costly and time-consuming enterprise. Cross-lingual learning methods try to bootstrap NLP tools for low-resource languages despite the lack of annotated resources in those languages. Early work focused on annotation projection or *data transfer* (Hwa et al. 2005; Yarowsky, Ngai & Wicentowski 2001). In that approach, annotations are

projected from the well-resourced source languages to the low-resource target language using parallel corpora. Secondly, in *model transfer*, models are trained on the annotated source language and applied to the target language. Without adaptation, this only works reasonably well for closely related languages (Agić et al. 2014). Typically, models need to be delexicalised unless there is a substantial lexical overlap between source and target language. Due to the availability of harmonised PoS annotations (Petrov, Das & McDonald 2012), shared PoS features across languages can be used by delexicalised models (McDonald et al. 2013), which can be further improved using cross-lingual word clusters or target language adaptation. Guo et al. (2015) extend delexicalized transfer models with cross-lingual distributed representations to include lexical knowledge. They apply alignment-based projection and canonical correlation analysis (CCA) to map monolingual distributed word representations. Lastly, translating treebanks is a cross-lingual method proposed by Tiedemann, Agić & Nivre (2014). A given source language treebank is translated by an existing MT system, for example a statistical MT model trained on parallel data. Annotations are projected from source to synthetic target language sentences with the same techniques as before with the main benefit that they come from manually verified annotations instead of automatically parsed data sets.

An overview of common techniques of cross-lingual methods is presented in (Tiedemann & Agić 2016). We base our work on a similar setup but move away from proxy-based evaluation to a real-world scenario. To the best of our knowledge, no other publications focus their work on truly under-resourced languages. Agić, Hovy & Søgaard (2015) come closest to this idea by focusing on languages with small treebanks using resources such as the Bible and the Watchtower corpus that cover many low-density languages. In contrast to previous work, we basically start from scratch, having a small treebank for development purposes with the goal of making the best use out of the tools and resources available for our target language, Maltese.

1.1 Maltese

Maltese is a language spoken by the people of the Maltese archipelago that lie in the Mediterranean Sea some 80 kilometres south of Sicily as well as several Maltese communities abroad, totalling around 450,000 speakers. The Maltese language had a very interesting development in that it was under the influence of many languages from different language families in the course of Malta's history and has therefore been classified as a mixed language (Aquilina 1958). Maltese is assumed to originate from an Arabic dialect, brought to Malta by the Arabic conquerors in 870, that is close to the dialects spoken by the inhabitants of Tunisia. Malta had very strong links with Sicily and Italy and the Christian world in general and lost contact with the Arabic community in the 19th century. The language shows significant signs of Romancisation. Not only the lexicon, which has over 55% elements of Italo-Romance origin, but also morphology, syntax, and semantics are influenced by Italo-Romance languages (Stolz 2011). English started to have an impact on the language during the 19th century and more notably during the 20th century, largely due to the bilingual situation. After Malta's independence in 1964, Maltese became an official language

and since its membership in the EU in 2006 it is also an official language of the union.

The previous paragraph underpins the statement that bootstrapping a dependency parser for Maltese is an excellent real-world test case for cross-lingual methods. The development of Maltese NLP tools and corpora happens to be at the exact point in time, where cross-lingual methods would be very useful. The first steps towards treebanking have been made: a few hundred sentences have been annotated. Just enough to evaluate the crosslingual methods, but not enough to train a good lexicalised parser. The Maltese NLP community is joining forces to develop NLP tools for Maltese but is behind the other European languages. As a result, some NLP tools (a PoS tagger and a chunker) have become available over the past couple of years and a large monolingual corpus has been released in 2016. An electronic dictionary with morphological information and translations into English has been released very recently as well. Fortunately for us, Malta's membership in the EU has resulted in the availability of reasonably large quantities of parallel data. Furthermore, the fact that Maltese is strongly influenced by a variety of languages and cultures makes it a fascinating case for multi-source transfer methods.

2 Cross-lingual parsing

In order to answer the question concerning the practicality of cross-lingual methods in a real-world scenario, we experimented with several flavours of cross-lingual parsing approaches that have proven successful in previous work. We will briefly introduce the approaches below.

2.1 Annotation projection

In this approach, we apply the heuristics proposed by Hwa et al. (2005) that follow the direct correspondence assumption to make it possible to map annotations from one language to another through automatic word alignment in bitexts. The projection rules ensure the creation of valid tree structures in the target, which is necessary for the training procedures we apply. To reduce the negative influence of dummy nodes, we add the improvements proposed by Tiedemann & Agić (2016). In particular, we rely on the removal of dummy leaf nodes and the conflation of unary productions in the parse tree that involve dummy nodes. Furthermore, we remove all sentences that include any dummy node or dummy relation in their parse tree after projection and the modifications mentioned above.

For each source language we then use 40,000 sentences with projected annotations to train target language parsers. We also introduce additional features to the projected data by looking up morphological features in a monolingual dictionary. PoS information is used for some basic disambiguation.

2.2 Model transfer

The simplest transfer approach is to use delexicalised models that rely entirely on universal PoS tags. These baseline models work surprisingly well on closely-related languages but have a lot of short-comings due to the over-simplification of their feature models and their strong structural correspondence assumptions. An interesting option for target language adaptation is relexicalization by means of self-training. The simple idea is to apply the delexicalized model to monolingual target language data to create automatically annotated data for training fully lexicalized parsing models.

2.3 Translating treebanks

A third cross-lingual parsing technique proposed by Tiedemann, Agić & Nivre (2014) applies machine translation to create synthetic training data with projected annotation from the original treebank. This approach has been shown to be quite successful and often better than annotation projection. One of the main advantages is the use of manually verified source language annotation instead of noisy parsed out-of-domain data.

The biggest disadvantage is, of course, the lack of translation quality. Another critical problem is the requirement of sufficient training data for creating MT models. Typically, we cannot expect to have sufficient parallel data available to train statistical MT models for low-resource languages. However, the situation is different for Maltese due to its status in the EU.

For the translation approach, we follow the basic setup of our previous work and train standard phrase-based SMT models for all language pairs using a standard pipeline that includes word alignment, phrase extraction and phrase scoring.[1] Word alignments are already available from the annotation projection experiments and use the same output of the symmetrised alignments produced by efmaral and its fertility-enhanced HMM model. Phrase extraction and phrase scoring use standard settings of the Moses pipeline (Koehn et al. 2007). The language model is estimated using kenlm (Heafield et al. 2013) with an order of 5 and modified Kneyser-Ney smoothing from the entire Maltese corpus described below.

3 Tools and resources

The **treebank** that we use as our test set was created by Slavomír Čéplö from Charles University in Prague. This corpus is still under heavy development. At the time of writing, 371 sentences have been tokenised and manually annotated with PoS tags and Universal Dependency (UD) relations by one annotator only. It contains sentences from the following domains: journalistic (239 sentences), short stories (14 sentences), encyclopedic and instructional (118 sentences).

[1] We skip tuning in our current experiments because there are no suitable development sets for Maltese. Nevertheless, standard weights are usually a good initial guess and the scores in our experiments suggest that the models produce reasonable output for the task at hand.

The **Korpus Malti v3.0**[2] is the latest achievement in an ongoing project of gathering digital resources for Maltese (Gatt & Čéplö 2013). It contains 250 million tokens in various genres. It is tagged with the Maltese tagger described below, with an accuracy of around 97%. In addition, it includes partial lemma information and morphological analysis. We use the corpus for our relexicalisation experiments and for language modeling in the translation approach.

Ġabra is a free, open lexicon for Maltese, built by collecting various different lexical resources into one common database. Ġabra was originally built in 2013 as part of a master's thesis project by Camilleri (2013) While it is not yet a complete dictionary for Maltese, recent efforts resulted in an on-line dictionary with 15,259 entries and 4,514,367 inflectional word forms. Many of these are linked by root, include translations in English, and are marked for various morphological features.

Tagging Maltese was carried out using the SVMTool (Giménez & Màrquez 2004), trained on a manually tagged corpus of ca. 28,000 tokens. The Maltese Tagset v3.0[3] was developed by Slavomír Čéplö and Albert Gatt.

The **parser** we apply in our experiments is a graph-based model implemented in the mate tools (Bohnet 2010) and we use version 3.61.[4] For training the source language models we used the treebanks provided by the Universal Dependencies (UD) project version 1.3.[5] Note that we reduce dependency relations to universal categories to make it possible to transfer labels across languages.

3.1 Parallel data sets

Our two main approaches, annotation projection and treebank translation, heavily rely on parallel data sets. The DGT translation memory (Steinberger et al. 2012) can be used to produce an aligned multilingual corpus of the European Union's legislative documents (Acquis Communautaire) in 24 EU languages. This memory contains most, although not all, of the documents which make up the Acquis Communautaire, as well as some other documents which are not part of the Acquis.

In our experiments we apply a subset of 19 languages for which we have sufficient data for all languages in parallel including Maltese. The final corpus we use for our experiments comprises in the end about 1.27 million truly parallel sentences with roughly 19 to 26 million tokens per language.

We tokenized Maltese using an in-house tokenizer that is compatible with the treebank test set and applied UDPipe[6] (Straka, Hajič & Straková 2016) for tokenising the other 18 languages according to the standards of the universal dependency treebanks. Thereafter, we aligned all languages with Maltese on the word level using efmaral,[7] an efficient implementation of fertility-based alignment models based on Gibbs sam-

[2] The corpus along with many other electronic resources for Maltese can be retrieved from `http://mlrs.research.um.edu.mt`.

[3] `http://mlrs.research.um.edu.mt/resources/malti03/tagset30.html`.

[4] `https://code.google.com/archive/p/mate-tools/downloads`.

[5] `http://universaldependencies.org/introduction.html`.

[6] `http://ufal.mff.cuni.cz/udpipe`.

[7] `https://github.com/robertostling/efmaral`.

pling with a Bayesian extension (Östling & Tiedemann 2016). The word alignments are then symmetrised using the intersection and the popular grow-diag-final-and heuristics. Both of these symmetrized alignments are facilitated in our annotation projection experiments according to the procedures proposed by Tiedemann (2015).

4 Results and discussions

In Table 1 we see the results of the three main methods presented in this paper. Scores are given for the Maltese test set described above with predicted PoS labels for a realistic estimation of performance. We can see that models trained on the translated treebanks are comparable but often better than the annotation projection results. Using re-lexicalisation of delexicalized models works surprisingly well for source languages like Spanish and Italian with attachment scores that are quite close to the corresponding annotation projection experiments. But in general, the scores are significantly lower. Note that the purely delexicalized models perform even worse.

Table 1: Results for the three methods and the 18 source languages.

	Projection		Relexicalisation		Translation	
src	LAS	UAS	LAS	UAS	LAS	UAS
bg	54.94	66.24	46.56	59.52	54.86	66.25
cs	53.54	64.44	44.58	56.04	53.73	65.21
da	52.87	63.77	41.43	51.56	51.54	62.31
de	54.33	65.05	40.30	49.94	55.58	65.45
el	36.79	53.02	34.22	44.01	37.87	56.27
en	59.39	**69.53**	51.11	62.14	59.62	68.88
es	**59.78**	69.41	55.54	**65.88**	**60.50**	**70.32**
et	35.02	48.42	28.48	44.97	35.82	52.53
fi	37.14	50.49	27.61	41.73	37.09	53.02
fr	57.73	67.64	53.09	63.46	58.70	68.65
hu	49.30	59.36	28.70	40.85	41.10	52.15
it	57.70	66.74	**56.04**	65.11	60.35	68.80
nl	51.23	59.71	41.84	49.89	51.38	60.48
pl	51.78	63.33	44.01	53.98	52.09	62.58
pt	54.04	63.74	50.34	58.85	55.20	65.17
ro	55.80	65.72	50.56	61.29	56.75	67.43
sl	58.46	67.59	45.39	55.34	57.90	67.58
sv	53.19	62.87	40.77	53.41	51.62	61.70

Is the fact that Maltese is a mixed language reflected in higher performances when transferring from languages that are known to have had an effect on the Maltese language? We can only give tentative answers due to the lack of in-depth analyses and comparable parallel data for Arabic. Nevertheless, the best languages for cross-

lingual learning are Italian and Spanish, two Romance languages, which is expected due to strong Romancisation of Maltese. A second place is taken by English. We obtain very good scores for projection and translation, but the re-lexicalised models lags behind. Linguistic literature acknowledges the influence of English, but states that it is more recent, and therefore less embedded in the language. This could explain why the re-lexicalised model, that relies for the larger part on the syntactic structure of the source language, falls behind. Looking at the results it seems possible to conduct comparative linguistic studies based on the success of cross-lingual learning. This would be an interesting avenue to explore in future work.

Another interesting question is whether multi-source models can be used to overcome individual weaknesses of the projected data sets. The mixed nature of Maltese may support a combination of source languages in particular. Table 2 shows that a simple concatenation of data from the several languages works surprisingly well. Unexpectedly, the linguistically biased combination of Romance languages and English does not lead to any gains over the model that uses all data sets. Another unbiased approach of selecting all source languages for which the supervised source language parsers reach a level of at least 80% LAS leads to only modest improvements over the combination of all languages, which further demonstrates the robustness of the simple multi-source approach. As a final step, we also tested the inclusion of inflectional features and lemmas coming from the lexical database of Maltese. To our disappointment, this leads to only minor improvements in LAS and even a slight drop in UAS. This suggests that adding lexical information without contextual disambiguation provides only little help but coverage issues may also be good reason for the failure of this approach.

Table 2: Multi-source projection models and combinations of methods.

Method	languages	LAS	UAS
Projection	all languages	62.51	71.54
Projection	en es fr it pt ro	62.52	71.28
Projection	bg cs en es it sl	62.77	**71.80**
Projection + inflectional info	bg cs en es it sl	**63.03**	71.54

4.1 Measuring the practicality of cross-lingual parsing

The numbers in the previous paragraphs, albeit interesting from a research perspective, have little practical value. In order to compare the merits of cross-lingual methods, we determined the amount of manually annotated data needed to reach levels of performance that are equal to those stemming from cross-lingual methods. We, therefore, split our small test set into tiny training data of various sizes while using the remaining examples for testing.[8] Figure 1 shows the learning curve in our

[8] In this setup, test data is always changing depending on how much of the data we reserve for training. Hence, these scores are not completely comparable but the general trends should still be trustworthy

procedure.

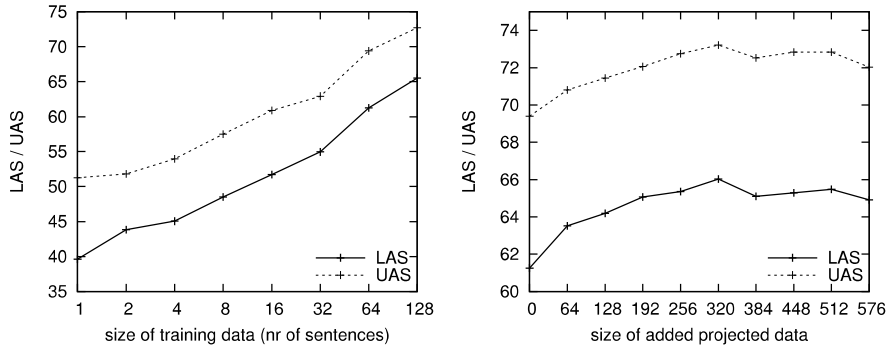

Figure 1: Figures showing a) the learning curve when training on manual data and b) the effect of adding projected data to a small amount of manually annotated data (64 sentences).

We can see that with as little as 64 training examples, we achieve a performance that is on par or above the best cross-lingual models discussed earlier. This comparison puts work on cross-lingual parsing into perspective. Most of the achievements presented in previous research are comparable to the results presented here but similar investigations on learning curves for tiny training data sets is usually not presented. The result above seems to strongly suggest that investing in annotation is a much wiser decision than spending time on tweaking a transfer model. A caveat may be that no expert can easily be found for many languages of the world and transfer models are still a valid choice for quickly building systems for a large number of languages. But the question remains whether this is really useful or not.

Another relevant question is whether hand-annotated data can successfully be combined with projected data to bootstrap models with very scarce resources. To study this, we ran another experiment in which we added small numbers of projected parse trees to a tiny treebank of 64 sentences for training parsing models that we then tested on the remaining test cases (with predicted PoS labels). The projected trees come from a multi-source model that we have re-trained on automatically parsed monolingual data from the Korpus Malti.

In this setup we use a simple concatenation strategy again and multiply the original treebank to match the size of the added noisy projected data. We can see in Figure 1 that the model trained on this augmented data indeed increases until a certain level of noise is reached that causes degradation of the parser. The improvements are still modest but it shows that there is some potential for such combined models. And given the small amount of projected trees necessary, there is room for filtering approaches to select high-quality projections.

to some extent.

5 Conclusions

We conducted a large number of experiments on dependency parsing to evaluate the practicality of cross-lingual learning for the real-world test case of Maltese, one of the official European languages, whose NLP tools and resources are under development. We leveraged all available resources in this realistic scenario and came to the following conclusions: firstly, the practicality of cross-lingual parsing is very much dependent on the current situation of the resources in the target language. Despite encouraging results with model and data transfer, cross-lingual parsing still lags far behind fully supervised models. For the scenario under discussion, small amounts of target language training data are available and these outperform state-of-the art cross-lingual models. In such a setting, a possible use of data transfer may be the combination with scarce manually annotated data sets to bootstrap treebanks and parsers. A scenario in which cross-lingual methods could be of practical use is that of a large number of languages for which no linguistic resources are available at all. However, the practical use of such rough models still remains to be proven. In addition to these observations regarding the practicality of cross-lingual parsing, we reflected on the possibility of using cross-lingual learning for comparative explorations regarding the similarity between languages on specific linguistic levels such as the syntactic level. We leave the large-scale, detailed analyses needed for such an endeavor to future work.

Acknowledgements

We would like to thank Slavomír Čéplö for making the manually annotated Maltese data available to us, as well as Albert Gatt for giving us access to the latest NLP tools and resources for Maltese.

References

Agić, Željko, Dirk Hovy & Anders Søgaard. 2015. If all you have is a bit of the Bible: learning POS taggers for truly low-resource languages. In *Proceedings of ACL*, 268–272.

Agić, Željko, Jörg Tiedemann, Danijela Merkler, Simon Krek, Kaja Dobrovoljc & Sara Može. 2014. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *Proceedings of LT4CloseLang*, 13–24.

Aquilina, Joseph. 1958. Maltese, a mixed language. *Journal of Semitic Studies* 3. 58–79.

Bohnet, Bernd. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING*, 89–97.

Camilleri, John. 2013. *A computational grammar and lexicon for Maltese.* Sweden: Chalmers University of Technology, Gothenburg MA thesis.

Gatt, Albert & Slavomír Čéplö. 2013. Digital corpora and other electronic resources for Maltese. In *Proceedings of the International Conference on Corpus Linguistics.* Lancaster, UK.

Giménez, Jesús & Lluís Màrquez. 2004. SVMTool: a general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th LREC*. Lisbon, Portugal.

Guo, Jiang, Wanxiang Che, David Yarowsky, Haifeng Wang & Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *In proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark & Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL*, 690–696.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas & Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering* 11(3). 311–325.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin & Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL*, 177–180.

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló & Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*, 92–97.

Östling, Robert & Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics (PBML)* (106). 125–146. http://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf.

Petrov, Slav, Dipanjan Das & Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*, 2089–2096.

Steinberger, Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos & Patrick Schlüter. 2012. DGT-TM: a freely available translation memory in 22 languages. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Stolz, Thomas. 2011. Maltese. In *The languages and linguistics of Europe: a comprehensive guide*, 241–256. De Gruyter Mouton.

Straka, Milan, Jan Hajič & Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA).

Tiedemann, Jörg. 2015. Improving the cross-lingual projection of syntactic dependencies. In *Proceedings of NoDaLiDa*.

Tiedemann, Jörg & Željko Agić. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research* 55. 209–248.

Tiedemann, Jörg, Željko Agić & Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of CoNLL*, 130–140.

Yarowsky, David, Grace Ngai & Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*, 1–8.