Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto (dir.)

Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015
3-4 December 2015, Trento

Accademia University Press

Gold standard vs. silver standard: the case of dependency parsing for Italian

Michele Filannino and Marilena Di Bari

DOI: 10.4000/books.aaccademia.1475
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: 11 November 2016
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788899200008

Gold standard vs. silver standard: the case of dependency parsing for Italian

Michele Filannino
The University of Manchester
School of Computer Science
M13 PL, Manchester (UK)
filannim@cs.man.ac.uk

Marilena Di Bari
University of Leeds
School of Languages, Cultures and Societies
LS2 9JT, Leeds (UK)
mlmdb@leeds.ac.uk

Abstract

English. Collecting and manually annotating gold standards in NLP has become so expensive that in the last years the question of whether we can satisfactorily replace them with automatically annotated data (silver standards) is arising more and more interest. We focus on the case of dependency parsing for Italian and we investigate whether such strategy is convenient and to what extent. Our experiments, conducted on very large sizes of silver data, show that quantity does not win over quality.

Italiano. *Raccogliere e annotare manualmente dati linguistici gold standard sta diventando oneroso al punto che, negli ultimi anni, la possibilita' di sostituirli con dati annotati automaticamente (silver) sta riscuotendo sempre piu' interesse. In questo articolo indaghiamo la convenienza di tale strategia nel caso dei dependency parser per l'italiano. Gli esperimenti, condotti su dati silver di grandissime dimensioni, dimostrano the la quantita' non vince sulla qualita'.*

1 Introduction

Collecting and manually annotating linguistic data (typically referred to as *gold* standard) is a very expensive activity, both in terms of time and effort (Tomanek et al., 2007). For this reason, in the last years the question of whether we can train good Natural Language Processing (NLP) models by using just automatically annotated data (called *silver* standard) is arising interest (Hahn et al., 2010; Chowdhury and Lavelli, 2011).

In this case, human annotations are replaced by those generated by pre-existing state-of-the-art systems. The annotations are then merged by using a committee approach specifically tailored on the data (Rebholz-Schuhmann et al., 2010a). The key advantage of such approach is the possibility to drastically reduce both time and effort, therefore generating considerably larger data sets in a fraction of the time. This is particularly true for text data in different fields such as temporal information extraction (Filannino et al., 2013), text chunking (Kang et al., 2012) and named entity recognition (Rebholz-Schuhmann et al., 2010b; Nothman et al., 2013) to cite just a few, and for non-textual data like in in medical imaging recognition (Langs et al., 2013).

In this paper we focus on the case of dependency parsing for the Italian language. Dependency parsers are systems that automatically generate the linguistic dependency structure of a given sentence (Nivre, 2005). An example is given in Figure 1 for the sentence "Essenziale per l'innesco delle reazioni è la presenza di radiazione solare." (The presence of solar radiation is essential for triggering the reactions). We investigate whether the use of very large silver standard corpora leads to train good dependency parsers, in order to address the following question: *Which characteristic is more important for a training set: quantity or quality?*

The paper is organised as follows: Section 2 presents some background works on dependency parsers for Italian; Section 3 presents the silver standard corpus used for the experiments and its linguistic features, with Section 4 describing the experimental settings and Section 5 describing the results of the comparison between the trained parsers (considering different sizes of data) and two test sets: gold and silver. Finally, the paper's contributions are summed up in Section 6.

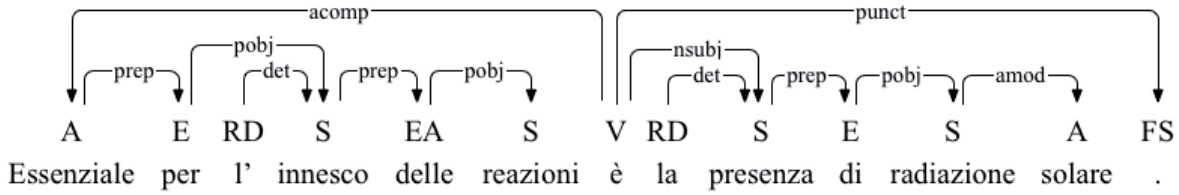

Figure 1: An example of dependency tree for an Italian sentence.

2 Background

Since dependency parsing systems play a pivotal role in NLP, their quality is crucial in fostering the development of novel applications. Nowadays dependency parsers are mostly data-driven, and mainly designed around machine learning classifiers. Such systems "train classifiers that predict the next action of a deterministic parser constructing unlabelled dependency structures" (Nivre, 2005).

Like in the case of other languages, in Italian ad-hoc cross-lingual and mono-lingual shared tasks are organised every year to push the boundaries of such technologies (Buchholz and Marsi, 2006; Bosco et al., 2009; Bosco and Mazzei, 2011; Bosco et al., 2014). The most important shared task about dependency parsing systems for Italian is hosted by the EVALITA series, in which participants are provided with manually annotated training data and the evaluation of their system is performed on a non disclosed portion of the data. Since the different systems presented so far have reached an overall performance close to 90% (Lavelli, 2014), we believe that the question of whether we can start using silver standards is a relevant one.

3 The corpus

The silver standard data comes from a freely available corpus created as part of the project PAISÀ (*Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati*) (Lyding et al., 2014). The project was aimed at "overcoming the technological barriers currently preventing web users from having interactive access to and use of large quantities of data of contemporary Italian to improve their language skills".

The PAISÀ corpus[1] is a set of about 380,000 Italian texts collected by systematically harvesting the web looking for frequent Italian collocations. It consists of about 13M sentences and 265M tokens fully annotated in CoNLL format. The average length of the sentences is about 20 tokens.

The Part-of-Speech tags have been automatically annotated by using ILC-POS-TAGGER (Dell'Orletta, 2009) and the dependency structure by using DeSR Dependency Parser (Attardi et al., 2007), the top performer system at the EVALITA shared task. The POS-tags are annotated according to the TANL tagset[2], whereas the dependency relations follow the ISST-TANL tagset[3]. These automatic annotations have been successively revised and manually corrected on different stages: text cleaning, annotation corrections and tools alignment.

Unfortunately we found out that The PAISÀ corpus includes some sentences which cannot be used for training purposes due to invalid CONLL representations (i.e. duplicated or missing IDs, and invalid dependency relations). These sentences represent the 6.04% of the corpus, yet only the 0.10% of the tokens. This difference shows the presence of many small invalid sentences.

Thus we have created a filtered corpus with the working sentences to which we will refer from now on with the name of *silver* as opposed to the EVALITA corpus as *gold*. In the latter, for training purposes we merged training and development test sets, whereas we did not modify the official test set.

4 Experiments

4.1 Test corpora

We quantitatively measured the performance of the proposed parsers with respect to two test sets: gold and silver.

[1] http://www.corpusitaliano.it/it/contents/description.html

[2] http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset

[3] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

	original	filtered	Δ%
Sentences	13.1M	12.3M	93.96%
Tokens	264.9M	264.6M	99.90%
Sentence length	20.3	21.5	-

Table 1: PAISÀ corpus' statistics. The figures show the presence of many short and invalid sentences.

The gold test set corresponds to the official benchmark test set for the EVALITA 2014 dependency parsing task. It contains 344 sentences manually annotated with 9066 tokens (∼26 tokens per sentence). The silver test set, instead, is composed of 1,000 randomly selected sentences from the silver data, which have not been used for training purposes in the experiments.

4.2 Experimental setting

The experiments have been carried out using eight different sizes of training set from the silver data: 500, 1K, 5K, 25K, 75K, 125K, 250K and 500K sentences. A limitation of the learning algorithm prevented us to consider even larger training sets[4].

We used the Unlabelled Attachment Score (UAS) measure which studies the structure of a dependency tree and assesses whether the output has the correct head and dependency arcs. The choice of UAS measure is justified by the fact that the gold and silver label sets are not compatible.

We trained the models with *MaltParser*[5] v.1.8.1 by using the default parameters.

The overall set of experiments took about a month with 16 CPU cores and 128Gb of RAM.

5 Results

The complete results are presented in Table 2. The 8 parsers trained on silver data perform poorly when tested against the gold test set (∼32%). The same happens for the opposite setting: the parser trained on the gold data and tested on the silver test set (last column of Table 2). By training on one set and testing on another (gold vs. silver), performance immediately drops of about 35%.

When the parser is trained on and tested against the gold data the performance is 85.85%. Such configuration corresponds to the EVALITA14 setting and provides results comparable with the one obtained by the afore-mentioned challenge's participants.

Training set		UAS against	
corpus	size	gold test	silver test
silver	500	30.14	66.11
	1.000	30.95	67.00
	5.000	32.21	69.11
	10.000	32.44	69.56
	25.000	32.83	69.92
	75.000	33.22	69.79
	125.000	33.47	70.27
	250.000	**33.58**	70.23
	500.000	33.20	**71.17**
gold	7.978	**85.85**	**48.30**

Table 2: Parsers' performance against silver and gold test sets. Silver data refers to PAISÀ corpus, whereas gold refers to EVALITA14 training and development set. Silver data have been used for training purposes in different sizes. Sizes are expressed in number of sentences.

The interesting result lies in the fact that providing a dataset 1000 times bigger does not significantly enhance the performance. This is true regardless of the type of test set used: gold (3.06% variance) and silver (4.89% variance). Moreover, training a parser on a data set smaller than its test set does not negatively affect the final performance.

Figure 2 depicts the performance curves for the models trained on silver data only.

In order to allow for the reproducibility of this research and the possibility of using these new resources, we make the dependency parsing models and the used data sets publicly available at `http://www.cs.man.ac.uk/~filannim/projects/dp_italian/`.

6 Conclusions

We presented a set of experiments to investigate the contribution of silver standards when used as substitution of gold standard data. Similar investigations are arising interesting in any NLP subcommunities due to the high cost of generating gold data.

The results presented in this paper highlight two important facts:

[4] The instance×feature matrix exceeds the maximum size allowed by the `liblinear` implementation used.

[5] `www.maltparser.org`

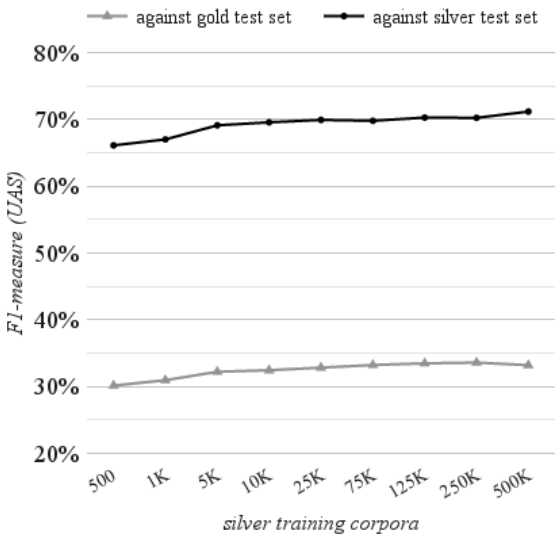

Figure 2: Parsers' performance against silver and gold test sets. In both cases, the models exhibit an asymptotic behaviour. Figures are presented in Table 2. Silver data sizes express the number of sentences. 'K' stands for 1.000.

- The size increase of the training corpus does not provide any sensible difference in terms of performance. In both test sets, a number of sentences between 5.000 and 10.000 seem to be enough to obtain a reliable training. We note that the size of the EVALITA training set lies in such boundary.

- The annotations between gold and silver corpora may be different. This is suggested by the fact that none of the parsers achieved a satisfactory performance when trained and tested on different sources.

We also note that the gold and silver test data sets have different characteristics (average sentence length, lexicon and type of annotation), which may partially justify the gap. On the other hand, the fact that a parser re-trained on annotations produced by a state-of-the-art system (DeSR) in the EVALITA task performs poorly on the very same gold set sheds light on the possibility that such official benchmark test set may not be representative enough.

The main limitation of this study lays in the fact that the experiments have not been repeated multiple times, therefore we have no information about the variance of the figures (UAS column in Table 2). On the other hand, the large size of the data sets involved and the absence of any outlier figure suggest that the overall trends should not change. With the computational facilities available to us for this research, a full analysis of that sort would have required years to be completed.

The results presented in the paper shed light on a recent research question about the employability of automatically annotated data. In the context of dependency parsing for Italian, we provided evidences to support the fact that the quality of the annotation is a far better characteristic to take into account when compared to quantity.

A similar study on languages other than Italian would constitute an interesting future work of the research hereby presented.

Acknowledgements

The authors would like to thank Maria Simi and Roberta Montefusco for providing the EVALITA14 gold standard set, and the two anonymous reviewers who contributed with their valuable feedback. MF would also like to thank the EPSRC for its support in the form of a doctoral training grant.

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Atanas Chanev, and Massimiliano Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using DeSR. In *EMNLP-CoNLLPAISA*, pages 1112–1118.

Cristina Bosco and Alessandro Mazzei. 2011. The EVALITA 2011 parsing task: the dependency track. *Working Notes of EVALITA*, 2011:24–25.

Cristina Bosco, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice DellâOrletta, and Alessandro Lenci. 2009. EVALITA'09 parsing task: comparing dependency parsers and treebanks. *Proceedings of EVALITA*, 9.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. *Proceedings of EVALITA*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

Faisal Mahbub Chowdhury and Alberto Lavelli. 2011. Assessing the practical usability of an automatically annotated corpus. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 101–109. Association for Computational Linguistics.

Felice Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. *Proceedings of EVALITA*, 9.

Michele Filannino, Gavin Brown, and Goran Nenadic. 2013. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 53–57, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Udo Hahn, Katrin Tomanek, Elena Beisswanger, and Erik Faessler. 2010. A proposal for a configurable silver standard. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 235–242. Association for Computational Linguistics.

Ning Kang, Erik M van Mulligen, and Jan A Kors. 2012. Training text chunkers on a silver standard corpus: can silver replace gold? *BMC bioinformatics*, 13(1):17.

Georg Langs, Allan Hanbury, Bjoern Menze, and Henning Müller. 2013. Visceral: Towards large data in medical imaging—challenges and directions. In *Medical Content-Based Retrieval for Clinical Decision Support*, pages 92–98. Springer.

Alberto Lavelli. 2014. Comparing state-of-the-art dependency parsers for the EVALITA 2014 dependency parsing task. *Proceedings of EVALITA*.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISA corpus of Italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43.

Joakim Nivre. 2005. Dependency grammar and dependency parsing. Technical report, Växjö University: School of Mathematics and Systems Engineering.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno-Yepes, Erik M van Mulligen, Ning Kang, Jan A Kors, David Milward, Peter T Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, et al. 2010a. The CALBC silver standard corpus for biomedical named entities-a study in harmonizing the contributions from four independent named entity taggers. In *LREC*.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M Van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010b. CALBC silver standard corpus. *Journal of bioinformatics and computational biology*, 8(01):163–179.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *EMNLP-CoNLL*, pages 486–495.