Building a Social Media Adapted PoS Tagger Using FlexTag – A Case Study on Italian Tweets

Tobias Horsmann **Torsten Zesch**
Language Technology Lab
Department of Computer Science and Applied Cognitive Science
University of Duisburg-Essen, Germany
{tobias.horsmann,torsten.zesch}@uni-due.de

Abstract

English. We present a detailed description of our submission to the PoSTWITA shared-task for PoS tagging of Italian social media text. We train a model based on FlexTag using only the provided training data and external resources like word clusters and a PoS dictionary which are build from publicly available Italian corpora. We find that this minimal adaptation strategy, which already worked well for German social media data, is also highly effective for Italian.

Italiano. *Vi presentiamo una descrizione dettagliata della nostra partecipazione al task di PoS tagging for Italian Social Media Texts (PoSTWITA). Abbiamo creato un modello basato su FlexTag utilizzando solo i dati forniti e alcune risorse esterne, come cluster di parole e un dizionario di PoS costruito da corpora italiani disponibili pubblicamente. Abbiamo scoperto che questa strategia di adattamento minimo, che ha già dato buoni risultati con i dati di social media in tedesco, è altamente efficace anche per l'Italiano.*

1 Introduction

In this paper, we describe our submission to the PoSTWITA Shared-Task 2016 that aims at building accurate PoS tagging models for Italian Twitter messages. We rely on FLEXTAG (Zesch and Horsmann, 2016), a flexible, general purpose PoS tagging architecture that can be easily adapted to new domains and languages. We re-use the configuration from Horsmann and Zesch (2015) that has been shown to be most effective for adapting a tagger to the social media domain. Besides training on the provided annotated data, it mainly relies on external resources like PoS dictionaries and word clusters that can be easily created from publicly available Italian corpora. The same configuration has been successfully applied for adapting FlexTag to German social media text (Horsmann and Zesch, 2016).

2 Experimental Setup

We use the FlexTag CRF classifier (Lafferty et al., 2001) using a context window of ± 1 tokens, the 750 most-frequent character ngrams over all bi, tri and four-grams and boolean features if a token contains a hyphen, period, comma, bracket, underscore, or number. We furthermore use boolean features for capturing whether a token is fully capitalized, a retweet, an url, a user mention, or a hashtag.

Data We train our tagging model only on the annotated data provided by the shared task organizers. As this training set is relatively large, we decided against adding additional annotated data from foreign domains which is a common strategy to offset small in-domain training sets (Ritter et al., 2011; Horsmann and Zesch, 2016).

Resources *Word clusters:* We create word clusters using Brown clustering (Brown et al., 1992) from 400 million tokens of Italian Twitter messages which have been crawled between the years 2011 and 2016.

PoS dictionary: We create a PoS dictionary which stores the three most frequent PoS tags of a word. We build the dictionary using a PoS annotated Italian Wikipedia corpus.[1]

Namelists: We furthermore use lists of first names obtained from Wikipedia and extract words tagged as named entities from the ItWaC web corpus (Baroni et al., 2009) to improve coverage of named entities.

[1] http://wacky.sslmit.unibo.it/doku.php?id=corpora

	Acc All	Acc OOV
TreeTagger Baseline	75.5	-
PoSTWITA	90.6	80.5
+ Clusters	92.7	85.6
+ PoS-Dict	92.2	85.3
+ Namelist	91.1	81.4
+ All Resources	**92.9**	**86.2**

Table 1: Results on the test data set

Baseline System We compare our results to the Italian model of TreeTagger (Schmid, 1995). As TreeTagger uses a much more fine-grained tagset than the one used in this shared-task, we map the fine tags mapping as provided by DKPro Core DKProCore (Eckart de Castilho and Gurevych, 2014).

3 Results

Table 1 gives an overview of our results. Besides the baseline, we show the results for only using the available training data (labeled *PoSTWITA*) and when adding the different types of external resources.

The baseline is not competitive to any of our system configurations, which confirms the generally poor performance of off-the-shelf PoS taggers on the social media domain. Using all resources yields our best result of 92.9%. Among the individual resources, word clusters perform best regarding overall accuracy as well as accuracy on out-of-vocabulary (OOV) tokens. This shows that clusters are also highly effective for Italian, as was previously shown for English (Owoputi et al., 2013) and German (Horsmann and Zesch, 2016).

We also computed the confidence interval by binomial normal approximation ($\alpha = 0.05$). We obtain an upper bound of 93.6 and a lower bound of 92.2. This shows that our best configuration is significantly better than using only the provided training data. Looking at the official PoSTWITA results, it also shows that there are no significant differences between the top-ranking systems.

Error Analysis In Table 2, we show the accuracy for each PoS tag on the test data set. The largest confusion class is between nouns and proper nouns, which is in line with previous findings for other languages (Horsmann and Zesch, 2016). It can be argued whether requiring the PoS tagger to make this kind of distinction is actually a good idea, as it often does not depend on syntactical properties, but on the wider usage context. Because of the high number of noun/proper confusions, it is also likely that improvements for this class will hide improvements on smaller classes that might be more important quality indicators for social media tagging. In our error analysis, we will thus focus on more interesting cases.

Tag	#	Acc	Primary Confusion
ADP_A	145	100.0	-
HASHTAG	115	100.0	-
MENTION	186	100.0	-
PUNCT	583	100.0	-
CONJ	123	99.2	VERB
URL	119	98.3	VERB
DET	306	95.8	PRON
ADP	351	95.7	ADV
PRON	327	93.3	DET
NUM	70	92.9	ADJ
INTJ	66	92.4	NOUN
NOUN	607	91.6	PROPN
VERB	568	91.6	AUX
AUX	109	90.8	VERB
ADV	321	90.3	SCONJ
SCONJ	60	90.0	PRON
ADJ	210	86.2	NOUN
EMO	79	83.5	SYM
PROPN	346	79.5	NOUN
VERB_CLIT	27	77.8	NOUN
SYM	12	72.7	PUNCT
X	27	55.6	EMO

Table 2: Accuracy per word class on the test data

In Table 3, we show examples of selected tagging errors. In case of the two adjective-determiner confusions both words occurred in the training data, but never as adjectives. The verb examples show cases where incorrectly tagging a verb as an auxiliary leads to a follow up error. We have to stress here that the feature set we use for training our PoS tagger does not use any linguistically knowledge about Italian. Thus, adding linguistically knowledge might help to better inform the tagger how to avoid such errors.

Amount of Training Data The amount of annotated social media text (120k tokens) in this

Adjective Confusions			
Token	Gold/Pred	Token	Gold/Pred
cazzo	INTJ	successo	VERB
sono	VERB	dal	ADP_A
tutti	DET	quel	ADJ / DET
sti	ADJ / DET	cazzo	NOUN
tweet	NOUN	di	ADP

Verb Confusions			
Token	Gold/Pred	Token	Gold/Pred
maggiormente	ADV	è	AUX / VERB
dell'	ADP_A	sempre	ADV
essere	VERB / AUX	stata	VERB / AUX
capito	ADJ / VERB	togliersi	VERB_CLIT
.	PUNCT	dai	ADP_A

Table 3: Adjective and Verb confusions

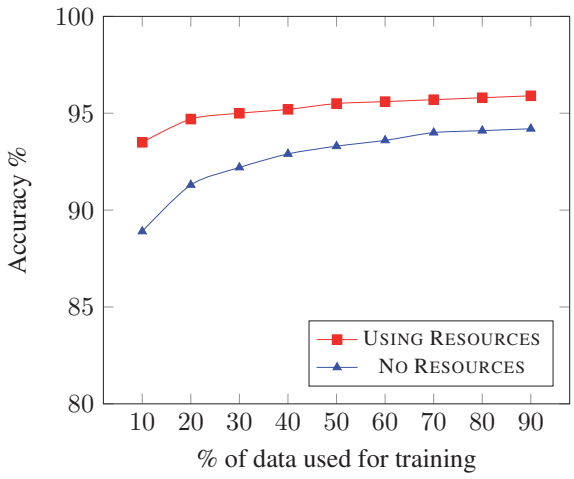

Figure 1: Learning Curve on training data with and without resources

shared-task is an order of magnitude larger than what was used in other shared tasks for tagging social media text. This raises the question of how much annotated training data is actually necessary to train a competitive social media PoS tagging model.

In Figure 1, we plot two learning curves that show how accuracy improves with an increasing amount of training data. We split the training data into ten chunks of equal size and add one additional data chunk in each iteration. We show two curves, one for just using the training data and one when additionally using all our resources. When using no resources, we see a rather steep and continuous increase of the learning curve which shows the challenges of the domain to provide sufficient training data. Using resources, this need of training data is compensated and only a small amount of training data is required to train a good model. The curves also show that the remaining problems are certainly not being solved by providing more training data.

4 Summary

We presented our contribution to the PoSTWITA shared task 2016 for PoS tagging of Italian social media text. We show that the same adaptation strategies that have been applied for English and German also lead to competitive results for Italian. Word clusters are the most effective resource and considerably help to reduce the problem of out-of-vocabulary tokens. In a learning curve experiment, we show that adding of more annotated data is not likely to provide further improvements and recommend instead to add more language specific knowledge. We make our experiments and resources publicly available.[2]

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Peter F Brown, Peter V DeSouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18:467–479.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland.

Tobias Horsmann and Torsten Zesch. 2015. Effectiveness of Domain Adaptation Approaches for Social Media PoS Tagging. In *Proceeding of the 2nd Italian Conference on Computational Linguistics*, pages 166–170, Trento, Italy.

[2] https://github.com/Horsmann/EvalitaPoSTWITA2016.git

Tobias Horsmann and Torsten Zesch. 2016. LTL-UDE @ EmpiriST 2015: Tokenization and PoS Tagging of Social Media Text. In *Proceedings of the 10th Web as Corpus Workshop*, pages 120–126, Berlin, Germany.

John D Lafferty, Andrew McCallum, and Fernando C N Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA.

Olutobi Owoputi, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Stroudsburg, PA, USA.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Torsten Zesch and Tobias Horsmann. 2016. FlexTag: A Highly Flexible Pos Tagging Framework. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4259–4263, Portorož, Slovenia.