Assigning part-of-speech to Dutch tweets

Mehdi Aminian, Tetske Avontuur, Zeynep Azar, Iris Balemans, Laura Elshof, Rose Newell, Nanne van Noord, Alexandros Ntavelos, Menno van Zaanen

Tilburg University
Tilburg, The Netherlands
{ M.M.AminianJazi, T.P.C.Avontuur, E.Z.Azar, I.J.P.Balemans, L.M.W.Elshof, R.K.Newell, N.J.E.vanNoord, A.Ntavelos, M.M.vanZaanen }@uvt.nl

Abstract

In this article we describe the development of a part-of-speech (POS) tagger for Dutch messages from the Twitter microblogging website. Initially we developed a POS tag set ourselves with the intention of building a corresponding tagger from scratch. However, it turned out that the output of Frog, an existing high-quality POS tagger for Dutch, is of such quality that we decided to develop a conversion tool that modifies the output of Frog. The conversion consists of retokenization and adding Twitter-specific tags. Frog annotates Dutch texts with the extensive D-Coi POS tag set, which is used in several corpus annotation projects in the Netherlands. We evaluated the resulting automatic annotation against a manually annotated sub-set of tweets. The annotation of tweets in this sub-set have a high inter-annotator agreement and our extension of Frog shows an accuracy of around 95%. The add-on conversion tool that adds Twitter-specific tags to the output of Frog will be made available to other users.

1. Introduction

Social media sites provide people with an easy and accessible forum to collaborate and share information. Social media can be grouped in six types: collaborative projects, blogs and microblogs, content communities, social networking sites, virtual game worlds, and virtual social worlds (Kaplan and Haenlein, 2010). These social media are extremely popular nowadays. For instance, Twitter generates approximately 200 million tweets (140-character messages) per day[1].

Given that social media generate so much data, it is interesting to investigate the potential of extracting useful information from the data being shared through these social media channels. In order to do so, some enabling technologies are essential. In the area of natural language processing, many tools rely on part-of-speech (POS) information. POS taggers (Voutilainen, 2003) assign tags that provide information on syntactic or morphological properties to words.

In this paper, we focus on the development of a POS tagger specifically for texts generated in a microblogging context. Microblogging services, such as Twitter, allow people to share information in the form of short messages. In the case of Twitter, a maximum of 140 characters are allowed per tweet or message. This small size has caused people to be very brief, sometimes even omitting words that may be obvious to human readers from the context.

The idea of developing a POS tagger for microblogging posts is based on the work by Gimpel et al. (2011), which describes the development of a POS tagger for English tweets. More information about this project can be found in section 2. Similarly to Gimpel et al. (2011), who worked on their project with 17 people, the project discussed in this paper has been accomplished by a group of students. More specifically, the group consisted of eight Master's students from Tilburg University who had just completed a Master's course in natural language processing. The authors not only come from varying scientific backgrounds (such as linguistics and computer science) but the group also had a variety of native tongues. In addition to the theoretical knowledge the students acquired during the natural language processing course, this project, which took approximately a week in person-hours, offered them a hands-on experience and insight into the practical decisions that need to be made when working on real-world natural language processing projects.

2. Background

This project is based on a similar project by Gimpel et al. (2011). They address the problem of POS tagging for English data from the microblogging service, Twitter. They develop a tag set, annotate data, develop features and conduct experiments to evaluate these features. The evaluation is designed in such a way to make it possible to test the efficacy of the feature set for POS tagging given limited training data. The features relate to Twitter orthography, frequently-capitalized tokens, the traditional tag dictionary, distributional similarity and phonetic normalization. The tagger with the full feature set leads to 89.37% accuracy on the test set. The project of Gimpel et al. (2011) was accomplished in 200 person-hours spread across 17 people and two months. With the results of their project, they want to provide richer text analysis of Twitter and related social media datasets. They also believe that the annotated data can be useful for research into domain adaptation and semi-supervised learning.

The effectiveness of the large amounts of data is shown in several studies. Keep in mind that while microblogging services generate large amounts of data, this also includes a large amount of "useless" data if one considers using the data for a particular purpose or when searching for information on a particular subject. Recently, there have been studies on the use of Twitter information in the area of sentiment analysis. In these cases, English POS tags are being used increasingly to analyze different aspects of social networks and Twitter in particular.

[1]`http://blog.twitter.com/2011/08/your-world-more-connected.html`

In the research paper, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", English POS tags for Twitter are used for the task of sentiment analysis (Pak and Paroubek, 2010). In this work, the researchers show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. They perform linguistic analysis of the collected corpus and explain the discovered phenomena. Using the corpus, they build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document. The researchers use TreeTagger for POS tagging and observe the difference in distributions of POS tags among positive, negative and neutral sets. Results show that some POS tags might be strong indicators of emotional text.

Other research that uses Twitter information focuses on a combination of sentiment and event detection on Twitter. An example of a study in this field is from Thelwall et al. (2011), who assess whether popular events are typically associated with increases in sentiment strength. They find strong evidence that popular events are normally associated with increases in negative sentiment strength and some evidence that the same is true for positive sentiment. However, the negative sentiment seems to be more central than the positive one.

Another example is the study about real-time event detection by social sensors by Sakaki et al. (2010). The authors devise a classifier of tweets based on features as keywords in a tweet, number of words, and their context. Secondly, they produce a probabilistic spatiotemporal model which can detect the center and the trajectory of the event location. In this work, every Twitter user is seen as a sensor. Filters are used to estimate location. Using this approach, they construct an earthquake reporting system in Japan. Because of the numerous earthquakes and Twitter users throughout Japan, they are able to detect earthquakes with a probability of 96%.

3. System overview

3.1. Tag set

Initially, we followed the process from Gimpel et al. (2011) in the development of a POS tag set for Dutch Twitter data. We started from their tag set and for each of the tags checked whether the tag made any sense in Dutch. It turned out that some of the English POS tags are not relevant in Dutch. For instance, the situation of nominal and verbal glued together, which is described by the 'L' POS tag, does not occur in Dutch.

To come up with better (non-Twitter specific) POS tags, we considered existing POS tag sets for Dutch, with the intention of extending these with Twitter specific tags. In this context, we looked at the POS tag set that is used in the SoNaR project.

SoNaR, which stands for Stevin Nederlandstalig Referentiecorpus is a corpus building project aiming at compiling a large corpus containing contemporary written Dutch (and Flemish). It is currently under development by Radboud University Nijmegen, Tilburg University, University of Twente, Utrecht University and KU Leuven. This project is financed within the Dutch-Flemish Stevin project[2] and is an extension of the D-Coi (Dutch Language Corpus Initiative) project (Oostdijk et al., 2008).

The tag set used in the SoNaR project is originally developed in the D-Coi project. The D-Coi tag set is described in more detail in Van Eynde (2005). This is an extensive tag set consisting of a total of 320 distinct tags. The tags are grouped by main tag, of which there are 13. Many specific tags are specializations of the main tag. For instance, the 'N' tag specifies nouns, which can be made more specific by adding arguments: 'N(soort,ev,basis,onz,stan)', which is a singular ('ev'), neuter ('onz'), common noun ('soort') in a non-diminutive ('basis') and nominal ('stan') form.

When analyzing the D-Coi tag set, it became clear that Twitter data requires some additional tags that are not present in the tag set used to annotate "regular" linguistic texts. Hence, we needed to extend the D-Coi POS tag set. When making decisions on which POS tags to select from the D-Coi tag set or to add, we took two factors into consideration:

1. the variety of parts-of-speech that can be encountered in Dutch tweets, and

2. the ease of user who will utilize our POS tagger.

In this sense we aim at choosing POS tags which will give enough information to discriminate POS of importance to the user.

Given the consideration and combining this with our aim for compatibility of the SoNaR project, we chose to base our tags on the main tags taken from the D-Coi tag set. The reason for not using the full D-Coi tag set is that we expected problems with manual annotation.

To incorporate Twitter-specific information, we had to add some Twitter-specific tags to the tag set. We chose to use the same Twitter-specific tags as Gimpel et al. (2011) used in their study. This led to the tag set that can be found in table 1. Two of these tags have more specific variants that deal with the more detailed linguistic aspects of the token it describes. 'N' has two sub-types: 'N(eigen)': proper nouns and 'N(soort)': common noun. 'SPEC' has seven sub-types to deal with symbols, incomprehensible words, abbreviations, etc.

However, during the development of the implementation of the POS tagger, we came across Frog[3], a POS tagger that can handle Dutch text and assigns tags according to the full D-Coi tag set. Initial experiments showed that the output of this tool is of such quality that it can also be used on Twitter data.

In the end, the availability of Frog, combined with the support of the POS tag set used in SoNaR and D-Coi allowed us to use the full D-Coi tag set. The only modification required was the addition of the Twitter-specific tags as shown in the rightmost column of table 1. These tags are based on the work of Gimpel et al. (2011).

[2] `http://lands.let.ru.nl/projects/SoNaR/`

[3] A more extensive description of the system is presented in section 3.2.

Table 1: Initial POS tag set.

Generic		Twitter	
ADJ	Adjective	AT	@ mention
BW	Adverb	DISC	Discourse marker
LET	Punctuation	EMO	Emoticon
LID	Determiner	HASH	# tag
N	Noun	URL	URL
SPEC	Special token		
TSW	Interjection		
TW	Number/ordinal		
VG	Conjunction		
VNW	Pronoun		
VZ	Preposition		
WW	Verb		

3.2. System implementation

The tweets were initially tagged using the POS tagger Frog, formerly known as Tadpole. Frog is a complete system that comes with the UCTO[4] tokenizer incorporated. Frog produces tab-delimited column-formatted output, one line per token. An example of such output can be found in table 2. The nine columns contain the following information (in order from left to right):

1. Token number (resets every sentence);
2. Token;
3. Lemma (according to the memory-based lemmatizer MBLEM[5]);
4. Morphological segmentation (according to the memory-based morphological analyzer MBMA[6]);
5. POS tag (D-Coi tag set; according to the memory-based tagger MBT[7]);
6. Confidence in the POS tag, which is a number between 0 and 1. This represents the probability mass assigned to the best guess tag in the tag distribution;
7. Chunker or shallow parser output on the basis of MBT;
8. Token number of head word (according to the constraint-satisfaction inference-based dependency parser, CSI-DP);
9. Type of dependency relation with head word.

From the Frog output we extract the token (2) and POS tag (5) columns and then automatically convert it to a Twitter-specific format. The conversion is based on a collection of regular expressions modifying the Frog output. This means that when needed we add Twitter-specific tags: 'HASH', 'AT', 'DISC', 'URL' or 'EMO'. In certain cases, this requires retokenization of the input. For instance, this is required when '#' or '@' tokens are found. In the cases of discourse markers or URLs, we changed the tag to DISC and URL respectively. As a URL we considered every token that begins with 'http://' or 'www.'. Moreover, URLs like 'http://www.youtube.com/watch?v=lRzFqW4Xh2k' which were separated by Frog at punctuation characters such as '=' in this case, are also retokenized.
Regarding the emoticons, we manually created a list of 156 emoticons that were found in the collection of tweets. We also included cases of big emoticons like: ':-))))))))'. Additionally, emoticons formed in a reversed fashion were added in the list because there are users that use emoticons in the way around (from right to left). This list covers the vast majority of the emoticons that are found in tweets.
Finally, the processing was done in parallel with the actual texts in order to avoid wrong conversion in cases similar to e.g. 'C# programming' which otherwise would lead to a tagging like '#programming' with a 'HASH' tag. Table 3 provides an example depicting the conversion of the Frog output to the Twitter-specific format. Note that the empty lines in the Twitter column are not in the output, but merely illustrate the alignment with the Frog column.

[4]http://ilk.uvt.nl/ucto/
[5]http://ilk.uvt.nl/mbma/
[6]http://ilk.uvt.nl/mbma/
[7]http://ilk.uvt.nl/mbt/

4. Experiments

To evaluate the quality of the output of the Frog POS tagger combined with the addition and modification of the output into Twitter-specific tags, we apply the tool to the collection of tweets that was provided by the SoNaR project. The output of this automatic annotation serves as an input for manual correction of the annotated tweets.
To perform the manual checking of the automatically annotated tweets we first tried to use the annotation tools Callisto[8] and MMAX2[9]. However, both systems turned out to be user unfriendly. Callisto cannot handle large amounts of tags (our POS tag set consists of 325 distinct tags). Changing tags using MMAX2 turned out to be difficult. In the end, we decided to use Gate[10]. Gate's annotation tool also had a minor disadvantage; it allows annotators to change the actual text (of the tweets), which is undesirable. Furthermore, it allows editing of the POS tags themselves, which can lead to inconsistencies.
We then evaluated the performance of the Twitter POS tagger by comparing the manually corrected output against the POS tagger output. In section 4.2., we provide information on the consistency of manual tagging/checking in the form of inter-annotator agreement and we will discuss the performance of the full system in the form of accuracy and F-score.

4.1. Dataset

The dataset that has been used in the experiments consists of 1,074,360 tweets. The large majority of these are tweets in Dutch, but we managed to identify a few non-Dutch tweets in the corpus. As mentioned earlier, the collection comes from the SoNaR corpus.
The original format of the tweets in the collection included among others timestamp, re-tweet information and any

[8]http://callisto.mitre.org/
[9]http://mmax2.sourceforge.net/
[10]http://gate.ac.uk/

Table 2: Frog column output.

1	Ze	ze	[ze]	VNW(pers,pron,stan,red,3,ev,fem)	1.000000	B-NP	2	su
2	vroeg	vragen	[vraag]	WW(pv,verl,ev)	0.532544	B-VP	0	ROOT
3	zich	zich	[zich]	VNW(refl,pron,obl,red,3,getal)	0.999740	B-NP	2	se
4	af	af	[af]	VZ(fin)	0.996853	O	2	svp
5	of	of	[of]	VG(onder)	0.733333	B-SBAR	2	vc
6	hij	hij	[hij]	VNW(pers,pron,nomin,vol,3,ev,masc)	0.999659	B-NP	8	su
7	nog	nog	[nog]	BW()	0.999930	B-ADVP	8	None
8	zou	zullen	[zal]	WW(pv,verl,ev)	0.999947	B-VP	5	body
9	komen	komen	[kom][en]	WW(inf,vrij,zonder)	0.861549	I-VP	8	vc
10	.	.	[.]	LET()	0.999956	O	9	punct

Table 3: Conversion from Frog to Twitter-specific output.

Frog		Twitter	
RT	SPEC(symb)	RT	DISC
@	SPEC(symb)		
nilicule	ADJ(prenom,basis,met-e,stan)	@nilicule	AT
#	SPEC(symb)		
sdgeld	WW(vd,vrij,zonder)	#sdgeld	HASH
http://t.co/74h22oo	SPEC(deeleigen)	http://t.co/74h22oo	URL
:	LET()		
-	LET()		
)	LET()		
)	LET()		
)	LET()	:-)))	EMO

URLs that are found in the tweet. In our project, we only considered the actual text of the tweets for further processing. All other information was discarded (but it is trivial to link the additional information with the POS tagged version of the tweets).

Going over the tweets manually, we identified specific aspects of the special nature of the tweets as texts in contrast to "regular text". Based on our qualitative analysis of Dutch tweets, we summarize those differences as follows:

Discourse markers Tweets may contain discourse markers like RT which is used when someone re-tweets another user's tweet. These types of discourse markers are typically not found in regular text.

@ mentions When a user wants to refer to another Twitter user, they use the character '@' before their Twitter user name;

 # tags People use the hash tag symbol '#' before relevant keywords in their tweet to categorize those tweets so that they are returned more easily as results of a Twitter search;

Alternative grammar and spelling Probably due to the limited length of a tweet (of at most 140 characters), tweets usually lack coherence. Also, they are sometimes written with limited grammar and non-perfect spelling.

An example of a typical Dutch tweet is: "RT @JoelSerphos: Kunnen de jongeren van #Iran de wereld net zo inspireren als hun leeftijdsgenoten in Egypte." (which translates to "RT @JoelSerphos: Can the youth of #Iran inspire the rest of the world just like their peers in Egypt." This tweet contains a discourse marker (RT), followed by an @ mention. Furthermore, "Iran" is used with a # tag.

4.2. Quantitative results

To conduct an evaluation of the generated output we need to build a gold standard dataset that contains POS tag annotation. We can then compare the output of the system against this gold standard dataset. For this purpose, we manually corrected the generated output of 1,056 tweets. This task has been done by three (Dutch) annotators who all manually corrected the POS tags of all of the approximately one thousand tweets.

To investigate the consistency with which the annotators agreed to the tags, we considered inter-annotator agreement. To measure inter-annotator agreement, we chose to use two measures: Cohen's Kappa and Fleiss' Kappa. Cohen's Kappa measures inter-annotator agreement between two annotators. Since we have three annotators, we compute this measure in pairs at a time, which leads to three results. We provide the pair-wise results in table 4 and also show the average inter-annotator agreement. Furthermore, to reach an overall inter-annotator agreement score, we also computed the Fleiss' Kappa which can compare multiple annotators at once. The results of the inter-annotator agreement can be found in table 4. As can be seen from this table, the inter-annotator agreement is very high. Note that the average Cohen's Kappa and the Fleiss' Kappa are only the same due to rounding.

Even though there may be some discussion on how to interpret these values, the inter-annotator agreement measures show consistently high values, which leads us to conclude

Table 4: Inter-annotator agreement of gold standard POS tags.

Measure	Annotators	Score
Cohen's Kappa	A vs. B	91.20
	A vs. C	92.07
	B vs. C	93.73
Average Cohen's Kappa		92.33
Fleiss' Kappa		92.33

Table 5: Evaluation results.

	Accuracy	F-Score
Complete tag set	92.87	92.61
Complete, simplified tags	94.12	93.94
Modified tokens	51.11	35.29
Modified tokens, simplified tags	50.57	34.43

that there was near complete agreement amongst the annotators. Note however, that the annotators corrected the POS tags and did not annotate the tweets from scratch, which would likely have led to a lower inter-annotator agreement.
During the process of manually correcting the POS tags of the tweets, the annotators noticed that the language used in the tweets corresponds highly with "regular" Dutch. As mentioned earlier, alternative spelling and grammar in tweets does occur, but not very frequently. Because of this, the quality of the output of Frog is already expected to be high. More research into the portion of creative use of language in tweets needs to be conducted to get a better idea on the impact of this phenomenon.
In table 5 the results of four evaluations are shown. Firstly, an evaluation is performed on the entire gold standard dataset with detailed POS tags (in other words, the full D-Coi tag set extended with the Twitter-specific tags). Secondly, the same evaluation is performed, but on a simplified tag set. For each of the complex POS tags, such as 'N(soort,ev,basis,zijd,stan)', only the main POS tag is used. In this case, the tag would be 'N'.
For the third and fourth evaluation only the tokens that are tagged differently by at least one of the annotators are taken into consideration. In table 5 these results are referred to as modified tokens. This comes down to 1,981 out of a total of 16,881 tokens in the gold standard dataset. Similarly to the first and second evaluation, the third evaluation makes use of the detailed tags while the fourth measures using the simplified tags.
Note that the modified tokens are the difficult tokens. The annotators did not necessarily agree in these cases. From the 1,981 tokens, there are 272 tokens for which the tags selected by the annotators did not lead to a majority vote. In these cases, a random selection was made.
Additionally, since the modified tokens are exactly the tokens where (at least one of) the annotators did not agree with the system output, we can expect that these results are much lower than the overall result. The fact that the accuracy for these cases is still around 50% means that the majority vote over all annotators still lead to the system output half of the time.
The results show that overall the performance of the POS tagger (Frog output converted into a Twitter-specific tag set) performs very well. Considering the complete (manually annotated gold standard) data set, accuracy and F-score are both over 90%.
Unfortunately, we cannot compare the output of our modified Frog against the output of plain Frog (without the conversion module). This is due to the retokenization of emoticons and URLs.

4.3. Qualitative results

During the manual annotation, the annotators encountered some consistent problems in the system output. URLs, for example, are hard to identify correctly because of tokenization problems. UCTO tokenizes parts of URLs, which leads to whitespace between parts of URLs, such as "echtbroodjeaap␣.␣nl". As a result of tokenization, the different tokens are annotated separately instead of as part of the URL.
Another difficulty is found with the tag that is used to annotate names ('SPEC(deeleigen)'). Although in most of the cases this tag is assigned to tokens correctly, sometimes this tag is too generic and a more specific tag would have been more appropriate. The tag set contains more specific tags for names, providing more information about the token such as gender, number, etc.
Another case deals with imperatives and interjections, which are also often tagged incorrectly. The latter, for example, is sometimes tagged as a verb instead of an interjection: "zeker", for instance, in the context of "ja, zeker!" (which translates to "yeah, sure"), is tagged as an adjective in its basic form. In this context, however, the token should obviously be tagged differently.
Finally, sometimes the system fails to recognize emoticons correctly. In some cases emoticons are not recognized where they should be recognized (false negatives). This is due to the fact that emoticons are used very creatively in tweets, which implies that a rather long list of emoticons is required in the system. In other cases, the system identifies an emoticon which is not a true emoticon (false positive). For instance, emoticons are found in places that do not practically allow for emoticons, such as within URLs, such as "(http://. . . =)".

5. Conclusion

Social media, Twitter in particular, is growing rapidly worldwide. In 2011 The Netherlands ranked #1 worldwide in penetration for Twitter users[11]. This rapid growth of Dutch tweets provides a great source of user-created contents in the Dutch language which can serve as an informal basis of information. However, to tap into this source of information, the data needs to be analyzed and understood. The POS tagger developed and presented in this paper can be applied to many linguistic analysis studies that involve Dutch tweets. This study provides a tool that enables a richer linguistic analysis of Dutch tweets.

[11] http://www.comscore.com/Press_Events/Press_Releases/2011/4/The_Netherlands_Ranks_number_one_Worldwide_in_Penetration_for_Twitter_and_LinkedIn

In this study we have modified the Frog POS tagger for Dutch to annotate Dutch tweets by adding a set of Twitter-specific tags. The results showed that it is possible to annotate Twitter-specific language. However, some problems remain. For instance, Frog finds it hard to identify URLs. This is partially solved by adjusting the conversion script, however, a modification of the UCTO tokenizer may be a more consistent solution. Furthermore, in this research we used a static list to recognize emoticons. This might pose a problem since emoticons are used creatively. A dynamic emoticon recognizer might help to deal with this creativity. Future work should include a deeper analysis of system errors and a possible modification of the conversion scripts to handle errors that are made consistently by the current system. Finally, to improve usability, the system should be build as a direct extension of Frog or perhaps even be included in the Frog distribution.

6. References

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers; Portland, OR, USA*, pages 42–47, New Brunswick:NJ, USA, June. Association for Computational Linguistics.

A.M. Kaplan and M. Haenlein. 2010. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, January–February.

N. Oostdijk, M. Reynaert, P. Monachesi, G. van Noord, R.J.F. Ordelman, I. Schuurman, and V. Vandeghinste. 2008. From d-coi to sonar: A reference corpus for dutch. In *Proceedings on the sixth international conference on language resources and evaluation (LREC 2008); Marrakech, Marokko*, pages 1437–1444. ELRA. ISBN=2-9517408-4-0.

A. Pak and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010); Valletta, Malta*, pages 1320–1326.

T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web; Raleigh, NC, USA*, pages 851–860. ACM, April.

M. Thelwall, K. Buckley, and G. Paltoglou. 2011. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

F. Van Eynde. 2005. Part of speech tagging en lemmatisering van het D-COI corpus.

A. Voutilainen. 2003. Part-of-speech tagging. In R. Mitkov, editor, *The Oxford handbook of computational linguistics*, pages 219–232. Oxford University Press, New York:NY, USA.