LANGUAGE MODEL BASED ON POS TAGGER

Bartosz Ziółko, Suresh Manandhar, Richard C. Wilson
Department of Computer Science, University of York, U.K.

Mariusz Ziółko
Department of Electronics, AGH University of Science and Technology, Kraków, Poland

Keywords: POS-tagging, language modelling, speech recognition, Polish.

Abstract: Language models are necessary for any large vocabulary speech recogniser. There are two main types of information which can be used to support modelling a language: syntactic and semantic. One of the ways to apply syntactic modelling is to use POS taggers. Morphological information can be statistically analysed to provide probability of a sequence of words using their POS tags. The results for Polish language modelling are presented.

1 INTRODUCTION

Part-of-speech (POS) (Brill, 1995) tagging is the process of marking up the words as corresponding to a particular part of speech, based on both its definition, as well as its context, using their relationship with other words in a phrase, sentence, or paragraph (Brill, 1995; Cozens, 1998). POS tagging is more than providing a list of words with their parts of speech, because many words represent more than one part of speech at different times. The first major corpus of English for computer analysis was the Brown Corpus (Kucera and Francis, 1967). It consists of about 1,000,000 words, made up of 500 samples from randomly chosen publications. In the mid 1980s, researchers in Europe began to use HMMs to disambiguate parts of speech, when working to tag the Lancaster-Oslo-Bergen Corpus (Johansson et al., 1978). HMMs involve counting cases and making a table of the probabilities of certain sequences. For example, once an article has been recognised, the next word is a noun with probability of 40%, an adjective with 40%, and a number with 20%. Markov Models are a common method for assigning POS tags. The methods already discussed involve operations on a pre-existing corpus to find tag probabilities. Unsupervised tagging is also possible by bootstraping. Those techniques use an untagged corpus for their training data and produce the tagset by induction. That is, they observe patterns in word structures, and provide POS types. These two categories can be further subdivided into rule-based, stochastic, and neural approaches. Some current major algorithms for POS tagging include the Viterbi algorithm, the Brill tagger (Brill, 1995), and the Baum-Welch algorithm (also known as the forward-backward algorithm). The HMM and visible Markov model taggers can both be implemented using the Viterbi algorithm.

POS tagging of Polish was started by governmental research institute IPI PAN. They created a relatively large corpus which is partly hand tagged and partly automatically tagged (Przepiórkowski, 2004; A.Przepiórkowski, 2006; Dębowski, 2003; Przepiórkowski and Woliński, 2003). The tagging was later improved by focusing on hand-written and automatically acquired rules rather than trigrams by Piasecki (Piasecki, 2006). The best and latest version of the tagger has accuracy 93.44%.

2 APPLYING POS TAGGERS FOR LANGUAGE MODELLING IN SPEECH RECOGNITION

There is little interest in using POS tags in ASR. Their usefulness was investigated. POS tags trigrams, a matrix grading possible neighbourhoods or probabilistic tagger can be created and used to predict a word being recognised based on left context analysed by a tagger. It is very difficult to provide tree structures, necessary for context-free grammars, which represent all possible sentences in case of Polish, as the order of words can vary significantly. Some POS tags are much more

Ziółko B., Manandhar S., C. Wilson R. and Ziółko M. (2008).
LANGUAGE MODEL BASED ON POS TAGGER.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 177-180
DOI: 10.5220/0001931701770180
Copyright © SciTePress

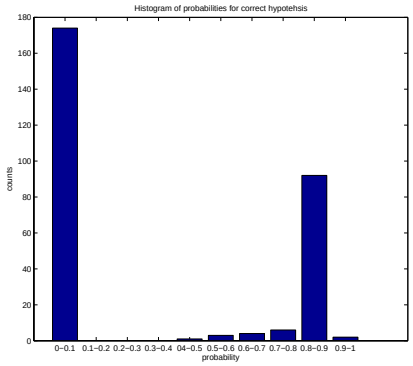

Figure 1: Histogram of POS tagger probabilities for hypotheses which are correct recognitions.

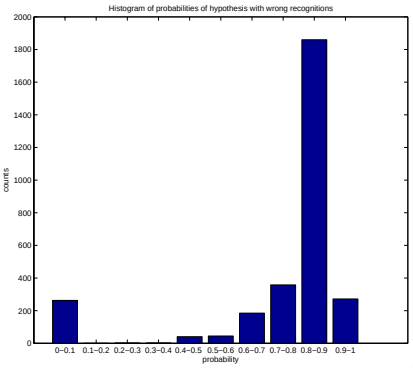

Figure 2: Histogram of POS tagger probabilities for hypotheses which are wrong recognitions.

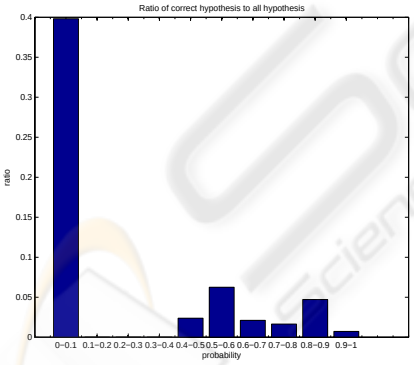

Figure 3: Ratio of correct recognitions to all for different probabilities from POS tagger.

probable in context of some others, which can be used in language modelling.

Experiments on applying morphological information to ASR of Polish language were undertaken using the best available POS tagger for Polish (Piasecki, 2006; Przepiórkowski, 2004). The results were unsatisfying, probably because to high ambiguity. An average word in Polish has two POS tags which gives too many possible combinations for a sentence. Briefly

Table 1: Results od applying the POS tagger to language modelling. First, a sentence in Polish is given, then a position of a correct recognition in a 10 best list. The description of tagger grade for the correct recognition follows.

Lubić czardaszowy pląs 1, Tagger grade is very low.
Cudzy brzuch i buzia w drzewie 4, Tagger grade is higher than for wrong recognitions.
Krociowych sum nie żal mi 1, Tagger grade is higher or similar then other recognitions in top 6 but lower then 7th.
Móc czuć każdy odczynnik 6, Tagger grade is lower than for most of the wrong recognitions including first two. However, the highest probability wrong recognition is grammatically correct.
On łom kładzie lampy i kołpak 7, Tagger grade is low.
On liczne taśmy w cuglach da 2, Tagger grade low but the highest in the first 5 hypoth.
Wór rur żelaznych ważył 3, Tagger grade is lower than for the first sentence.
Boś cały w wiśniowym soku 3, Tagger grade is higher then for 7 top hypotheses.
Lech być podlejszym chce 1, Tagger grade is the lowest in top 5 hypotheses but most of them are grammatically correct.
Żre jeż zioła jak dżem John 1, Tagger grade is higher than for top 4 hypotheses.
Masz dzisiaj różyczkę zieloną 1, Tagger grade is lower than for the second hypothesis which has no sense but morphologicaly is correct.
Weź daj im soli drogi dyzmo 2, Tagger grade is very close to the most probable hypothesis, which is grammatically correct.
Weź masz ramki opolskie 1, Tagger grade is higher than for the second hypothesis but lower than for the third one.
Dźgnął nas cicho pod zamkiem 1, Tagger grade is highest of all.
Tam śpi wojsko z bronią 6, Tagger grade is second, the highest one is 5th.
Nie odchodź bo żona idzie 3, Tagger grade is highest but equal to three others, which has acoustical probability lower.
Tym można atakować 5, Tagger grade is higher than for the acoustically most probable sentence but lower than for all other between 1 and 5, however all of them are grammatically correct.
Zmyślny kot psotny ujdzie 1, Tagger grade is higher then 2nd and 3rd hypothesis.
Niech pan sunie na wschód 4, Tagger grade is higher than for 7 most probable.

Table 2: Results od applying the POS tagger on its training corpus. First version of a sentence is a correct one, second is a recognition using just HTK and third one using HTK and POS tagging.

htk is better
skinęła głową zawstydzona
skinęła głową zawstydzona
skinęła bo w w w zawstydzona
poleż teraz spokojnie
poleż teraz spokojnie
poleż z teraz spokojnie
pamiętasz że opuściła sanktuarium
pamięta że opuściła sanktuarium
pamięta że opuściła sanktuarium w
o tak pamiętała wszystko powróciło z pełną wyrazistością
o tak pamięta wszystko powróciło pełną wyrazistością
o tak w pamięta wszystko powróciło pełną wyrazistością
same
nie mówiąc o tym kim ja jestem
nie w wiem nocy nocy nie jestem
nie w wiem nocy nocy nie jestem
zasłonię okno bo widzę że światło cię razi
zasłonię o okno bo widzę cię światło cię razi
zasłonię o okno bo widzę cię światło cię razi
zobaczysz wszystko będzie dobrze
zobaczysz wszystko będzie dobrze
zobaczysz wszystko będzie dobrze

speaking applying POS tagging for modelling of Polish is a process of guessing based on uncertain information.

HTK (Young, 1996; Young et al., 2005) was used to provide 10 best list of acoustic hypotheses for sentences from CORPORA. This model was trained in a way which allowed all possible combinations of all words in a dictionary. Then probabilities of those hypotheses using the POS tagger (Piasecki, 2006) were calculated. Acoustic model can be easily combined with language models using Bayes' rule by multiplying both probabilities.

3 EXPERIMENTAL RESULTS

Our experiments were conducted applying HTK on a corpus called CORPORA, created under supervision of Stefan Grocholewski in Institute of Computer Science, Poznań University of Technology in 1997 (Grocholewski, 1995). Speech files in CORPORA were recorded with the sampling frequency $f_0 = 16$ kHz equivalent to sampling period $t_0 = 62.5$ μs. Speech was recorded in an office with a working computer in the background which makes the corpus not perfectly clean. SNR (Signal to Noise Ratio) is not stated in the description of the corpus. It can be assumed that SNR is very high for actual speech but minor noise is detectable for periods of silence. The database contains 365 utterances (33 single letters, 10 digits, 200 names, 8 short computer commands and 114 simple sentences), each spoken by 11 females, 28 males and 6 children (45 people), giving 16425 utterances in total. One set spoken by a male and one by a female were hand segmented. The rest were segmented by a dynamic programming algorithm which was trained on hand segmented ones. None of the CORPORA utterances were in the original set used during development.

Trigrams of tags were calculated using transcriptions of spoken language and existing tagging tools. Results were saved in XML.

The results were compared giving different weights for probabilities from HTK acoustic model and POS tagger language model. In all situations the outcome probability gave worse results then pure HTK model. Histograms of probabilities for correct and wrong recognition were also calculated and they showed unuseful correlation. Some examples of sentences were also analysed and described by human supervisor. They are presented in Table 1.

In total 331 occurrences were analysed. Only 282 of them had correct recognition in the whole 10 best list of a given utterence. An average HTK probability of correct sentences was 0.1105. Exactly 244 of all occurrences had a correct hypothesis on the 1 position of the 10 best list. 0.7372 % of occurrences were correctly recognised while using only HTK acoustic model. Only 53 occurrences were recognised applying probabilities from the POS tagger, even when HTK probabilities were 4 times more important than those from POS tagger. The weight was applied by raising HTK probability to power of 4. It gives 0.1601 % of correct recognitions for a model with POS tag probabilities, which is a disapointing result.

Another way of proving usefulness of a model is through calculating histograms p_{posc} of probabilities received from the tagger for hypotheses which are correct recognitions (Fig. 1) and histogram p_{posw} of probabilities received from semantic model for hypotheses which are wrong recognitions (Fig. 2). The ratio $p_{posc}/(p_{posc} + p_{posw})$ is presented in Fig. 3. It shows that there is no correlation between high probability from the tagger and correctness of recognition.

The POS tagger was trained on a different corpus than the one used in an experiment described above. This is why we decided to conduct an additional experiment. We recorded 11 sentences from the POS tagger training corpus. They were recognised

by HTK, providing 10 best list and used in a similar experiment as the one described above. The amount of data is not enough to provide statistical results but observations on exact sentences (Table 2) provide the same conclusion as in the main experiment. The recognitions, which were found using HTK only, had fewer errors for 6 sentences. then 5 times the number of errors was the same. One sentence was correctly recognised for both models. One more was correctly recognised using just HTK acoustic model.

4 RECOGNITION USING LANGUAGE MODEL

Recognition can be conducted by finding the most coherent set of POS tags in a provided hypothesis. The tagger calculates P_{pos}, which can be used as additional weight in providing speech recognition due to Bayes' theorem. The values of p_{htk} probability gained from HTK model tend to be very similar for all hypotheses in the 10 best list of a particular utterance. This is why an extra weighting w was introduced to favour probabilities from audio model over p_{pos} received from the tagger. The final measure can be obtained applying Bayes' rule

$$p = p_{htk}^{w} p_{pos}. \quad (1)$$

Bayes rule is often used to compute posterior probabilities given observations. It can be used to compute the probability that a proposed hypothesis is correct, given an observation. It is often applied to combine probabilities of different models. p_{htk} is a probability of acoustic units given a word and p_{pos} is a probability of word. There should division by a probability of acoustic for normalisation purposes. It can be skipped as long as we deliver normalisation in another way or we accept the fact that final result is not a probability function, as it does not take values from 0 to 1. We can easily accept it if we are interested only in argument of a maximum of the result and we do not need proper probability values. Applying some linguistical data in speech recognition is necessary because acoustic models are not effective enough. However, the model based on POS tagger seems to not solve the issue.

5 CONCLUSIONS

It seems that POS tags are too ambiguous to be used effectively in modelling Polish for ASR. Another source of linguistical data has to be used to provide effective language model.

ACKNOWLEDGEMENTS

We received a significant help from Maciej Piasecki from the Technical University of Wrocław by providing tagger output and from Stefan Grocholewski from technical University of Poznan by letting us experiment on CORPORA.

REFERENCES

A.Przepiórkowski (2006). The potential of the IPI PAN corpus. *Poznań Studies in Contemporary Linguistics*, 41:31–48.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, December:543–565.

Cozens, S. (1998). Primitive part-of-speech tagging using word length and sentential structure. *Computaion and Language*.

Dębowski, Ł. (2003). A reconfigurable stochastic tagger for languages with complex tag structure. *The Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL*.

Grocholewski, S. (1995). Założenia akustycznej bazy danych dla języka polskiego na nośniku cd rom (eng. Assumptions of acoustic database for Polish language). *Mat. I KK: Głosowa komunikacja człowiek-komputer, Wrocław*, pages 177–180.

Johansson, S., Leech, G., and Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Olso/Bergen Corpus of British English, for Use with Digital Computers.* Department of English, University of Oslo.

Kucera, H. and Francis, W. (1967). *Computational Analysis of Present Day American English*. Brown University Press Providence.

Piasecki, M. (2006). Hand-written and automatically extracted rules for polish tagger. *Lecture Notes in Artificial Intelligence, Springer*, W P. Sojka, I. Kopecek, K. Pala, eds. Proceedings of Text, Speech, Dialogue 2006:205–212.

Przepiórkowski, A. (2004). *The IPI PAN Corpus: Preliminary version*. IPI PAN.

Przepiórkowski, A. and Woliński, M. (2003). The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. *Proceedings of the* 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, EACL 2003*.

Young, S. (1996). Large vocabulary continuous speech recognition: a review. *IEEE Signal Processing Magazine*, 13(5):45–57.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2005). *HTK Book*. Cambridge University Engineering Department, UK.