Semantic Modelling for Speech Recognition

Bartosz Ziółko,* Suresh Manandhar,*
Richard C. Wilson,* and Mariusz Ziółko**

*Department of Computer Science, University of York, UK
**Department of Electronics AGH, Poland
{bziolko,suresh,wilson}@cs.york.ac.uk
ziolko@agh.edu.pl

ABSTRACT

A new method of semantic modelling for speech recognition is presented. The method has some similarities to latent semantic analysis, but it gave better experimental results, which are provided as percentage of correctly recognised sentences from a corpus. The main difference is a choice of similar topics influencing a matrix describing probability of words appearing in topics.

1. Introduction

Semantic analysis is an important innovation of automatic speech recognition (ASR) as the very last step of the process. It can be applied as an additional measure to use the non-first choice word recognition hypothesis if they do not fit semantic content.

It is not efficient to recognise speech using acoustic information only. Human perception system is based on catching context, structure and understanding combined with recognition procedure. It is much easier to recognise and repeat without any errors a heard sentence if it is in a language we understand, comparing to a sentence in a language we are not familiar with. Language modelling can improve recognition highly. Semantic analysis can be done in many different ways and has been applied in ASR already. However, this kind of modelling is difficult due to the data sparsity problem. The literature always mentions semantic analysis as a necessary step in ASR but it is very difficult to find any research papers which provides any exact results on recognition when applying semantic methods.

Latent semantic analysis (LSA) [1] is a technique in natural language processing patented in 1988. It assumes that the meaning of a small part of text, like paragraph or sentence, can be approximated by the sum of the meanings of its words. LSA uses a word-paragraph matrix which describes the occurrences of words in topics. It is a sparse matrix whose rows correspond to topics and columns correspond typically to words that appear in the topics. The elements of the matrix are proportional to the number of times the words appear in each document, where rare words are upweighted to reflect their relative importance. This process is conducted by singular value decomposition (SVD). There are other methods of analysing semantic information, like topic signatures [2, 3]. Our method is the most closely connected to LSA.

Recently, graph based methods [4–6] are more and more popular. We used a graph instead of applying SVD to smooth information between different topics. Graphs help us to locate and grade similar topics.

2. Experimental Setup

Semantic analysis might be much more crucial in non-positional languages than in English due to irregularities in position structures of words. Language models based on context free grammars are quite unsuccessful for non-positional languages. Research about applying LSA in ASR was done [1] for English only, according to our knowledge.

English is the most common language of speech recognition research with Chinese and Japanese as two other common languages. This paper is focused on ASR of Polish. There is quite little research and no working continuous Polish speech recognition system. Polish and English are languages of the same Indo-European group, but there are some important differences between these languages which have an impact on ASR:

- English has a large number of homophones. What is more, many combinations of different words have similar pronunciation. Polish has fewer homophones.
- Pronunciation of vowels in English is very similar. If a vowel is not stressed it is pronounced as /ə/ or /I/. What is more, both of these phonemes have quite similar sounds and spectra. It means that unstressed vowels are almost indistinguishable in English. It contrasts with Polish.
- Modern English has emerged as a mixture of around thirty languages. It resulted in quite simple general rules (which was necessary for a language to be widely accepted by different people) but many irregularities (as a kind of residues), especially in pronunciation. Modern Polish is strongly based on Latin. In contrary to English, it resulted in very complicated grammar rules and morphology but quite few irregularities in pronunciation.
- English is a positional language, while Polish is an inflective one. A meaning of a word in English depends strongly on a position of a word in a sentence. In Polish a position has a secondary importance, an exact meaning of a word depends mainly on morphology. This fact means that the usage of syntax modelling is very difficult for Polish and possibly not as necessary as for English.
- In English, conjugation and declension are relatively simple and adjectives do not need any type of agreement. In Polish there are groups of different ways of conjugation and declension. Adjectives and numbers are agreed with nouns they describe. There is no general rule of word agreement. Different groups of words have their own types of endings. Verbs have 47 inflection forms (excluding participles), adjectives 44, numerals up to 49, adverbs 3, nouns and pronouns 14.

The method was tested on a set of Polish sentences from CORPORA [7]. Speech files in CORPORA were recorded with the sampling frequency f_0 = 16 kHz in an office with a working computer in the background. Signal to noise ratio (SNR) 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 con-

tains 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.

HMM Toolkit (HTK) [8, 9] was used to provide 10 best lists of acoustic hypotheses for sentences from the corpus. The audio model was trained on the complete CORPORA but in the semantic experiment we used only the mentioned 114 simple sentences as all other utterances are obviously too short to use them in language modelling. The mel-frequency cepstral coefficients (MFCC) [10, 8] were calculated for parameterisation. 12 MFCCs plus an energy with first and second derivatives were used, giving a standard set of 39 elements. We used 25 ms windows for audio framing and preemphasis filtering 0.97. Segments were windowed using Hamming method. 37 different phonemes were distinguished using a phonetic transcription provided with the corpus.

3. Training Algorithm

The entire algorithm was illustrated on a simple English example in one of the following sections. The corpus was organised as follows: all letters are combined in one topic, all digits in another, names and commands separately in two more. Every sentence is treated as a topic. In this way 118 topics are provided. They all consist of 659 different words in total. Then a matrix

$$S = [s_{ik}] \tag{1}$$

representing semantic relations is created, where rows $i = 1, ..., I$ represent topics and columns $k = 1, ..., K$ represent words. Each matrix value s_{ik} is the number of times word k occurs in topic i. A measure of similarity between two topics is

$$d_{ij} = \sum_{k=1}^{K} S_{ik} S_{jk} \tag{2}$$

It has to be normalised according to formula

$$d'_{ij} = d_{ij} / \max_{i<j} \{d_{ij}\}. \tag{3}$$

As a result we obtain $0 \le d'_{ij} \le 1$.

These topic similarities were analysed as follows:

1. Create an undirected, complete graph (Fig. 1) with topics as nodes and d'_{ij} as weights of edges. Let us define a path weight

$$p_{ij} = \prod_{(a,\,b) \in P(i,\,j)} d'_{ab}\,, \tag{4}$$

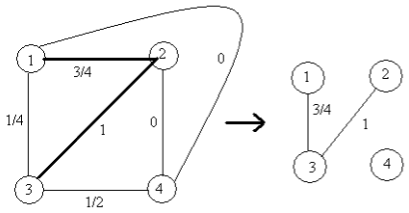

Figure 1. Undirected, complete graph illustrating similarities between sentence.

where $P(i, j)$ is the sequence of edges in the path from i to j. In the simplest case of a single edge i to j path weight is d'_{ij}. In case of multiple edges path, it is a product of similarities of all edges on a path (4).

2. For each node, we need to find n nodes with highest measures of paths leading to them from the given node. That will allow us to define a list N of semantically related topics which consists of the n nodes with their measures. The exact implementation of this part is presented in the next section.
3. The matrix S has to be recalculated to include impact of similar topics. It is expressed by a matrix

$$S' = [s'_{ik}] \; . \tag{5}$$

For all topics in matrix (1), we add all values of topics from the list of related topics, multiplied by a measure for a given pair of topics. The elements of S' are

$$s'_{ik} = s_{ik} + \alpha^{-1} \sum_{j \in N} p_{ij} \, s_{jk} \; . \tag{6}$$

Coefficient α is a smoothing factor which provides additional weight for influence of other topics on matrix S'.

4. Process of Finding Most Similar Topics

The group of the longest paths, where a distance is calculated using product between edges rather than sum has to be found in the 2nd point of the algorithm described in the previous section. It can be achieved by implementing the following algorithm:

1. Find n single edge paths with the highest measures d'_{ij}.
2. Check if the two edges path $P(i, m)$ starting from the node i with the highest measure d'_{ij}, which was found in the step above and going through j to any other edge m, has a better measure p_{im} than the lowest of the n solutions found in the step above. If it does than replace it with m in the list of n similar topics.
3. Conduct the step above for all other single node paths from the list apart from the lowest, nth element.
4. If there are any non single edge paths $P(i, j)$ on the list on position different then nth, repeat a process similar to step 2. Check if after adding any other edge a measure of path p_{ij} is higher than a measure of the nth position. Than replace the previous path with a new longer one path with higher p_{ij}.

It can be proved that the process is exhaustive in one way (from the analysed topic). Let us name the analysed topic as i and the set of the n most similar topics found in the first step of the process (using a measure d'_{ij}) to i as N_1. Let l be the element with the lowest measure of similarity d'_{ij} of N_1. As a result of the algorithm presented above we obtain

$$d'_{in_1} > d'_{ij} \quad \forall\, n_1 \in N_1, \forall\, j \notin N_1 \tag{7}$$

$$d'_{in_1} \geq d'_{il} \quad \forall\, n_1 \in N_1 \tag{8}$$

Let us define a set $N_2 = T / (\{i_a\} \cup N_1)$ of topics not included in the list of similar topics, where T is a set of all topics and $\{i_a\}$ is a one element set with the analysed topic i_a. From definition (3)

$$0 \leq d'_{ij} \leq 1 \quad \forall\, i,j \in \{1, ..., I\}\ , \tag{9}$$

therfore

$$d'_{in_1}\ d'_{n_1 j} \leq d'_{in_1} \quad \forall\, n_1 \in N_1, \forall\, j \in N_2\ . \tag{10}$$

It is also true that

$$d'_{in_1} \leq d'_{ij} \quad \forall\, j \in N_2 \tag{11}$$

Therefore

$$d'_{in_1}\, d'_{n_1 j} \leq d'_{ij} \quad \forall\, j \in N_2\ . \tag{12}$$

As the same reasoning can be applied for further iterations (three edges paths and so on) (10) and (12) prove that the process is exhaustive in one way. It can skip some solutions from other topics to the analysed one but it is better from linguistical point of view as we do not want topics assigned as being similar to many other topics just because they have a very strong link to one other topic.

5. Example in English

Let us consider an example of a corpus consisting of four sentences, all of them are treated as separate topics. *Big John has a house. Big John has a black, aggressive cat. The black aggressive cat has a small mouse. The small mouse is a mammal.*

All articles were skipped as they have no semantic content and they do not exist in Polish which was our experimental language. We count all other words creating a matrix S (Tab. 1), which gives following topic similarities ($d'_{12} = 3/4$, $d'_{13} = 1/4$, $d'_{14} = 0$, $d'_{23} = 1$, $d'_{24} = 0$, $d'_{34} = 1/2$) constructing the graph (Fig. 1) from matrix D (Tab. 2). We search for $n = 2$ similar topics. Applying first step of the process on the graph $N_1 = \{2, 4\}$ is found, where topic 4 is l, the topic with the lowest measure in N_1, namely 1/2. In the next step, p_{ij} are calculated for two edges paths starting at node 3 and going through 2. There are two of them. First one is for the path 3-2-4, where $p_{34} = 1 \bullet 0 = 0$. The second one is for the path 3-2-1, where $p_{31} = 1 \bullet 3/4 = 3/4 > d'_{34}$. This is why the topic 4 is replaced with topic 1 and the final list of topics similar to 3 is $\{2, 1\}$. Then assuming $\alpha = 2$ we can calculate a row for topic 3 from S' (Tab. 1).

Table 1. Matrix S for the example and one row of S' for the topic 3

topic	big	John	has	house	black	aggr.	cat	small	mouse	is	mam.
1	1	1	1	1	0	0	0	0	0	0	0
2	1	1	1	0	1	1	1	0	0	0	0
3	0	0	1	0	1	1	1	1	1	0	0
4	0	0	0	0	0	0	0	1	1	1	1
3'	**7/8**	**7/8**	**15/8**	**1/2**	**11/8**	**11/8**	**11/8**	**1**	**1**	**0**	**0**

Table 2. Matrix D for the example

	1	**2**	**3**	**4**
1	4	3	1	0
2	3	6	4	0
3	1	4	6	2
4	0	0	2	4

6. Recognition Using Semantic Model

Recognition can be conducted by finding the most coherent topic for a set of words W in a provided hypothesis. It is carried on by finding a maximum of a sum of elements of (5) from columns representing the words from a hypothesis over rows

$$p_{sem} = \max_{i \in W} \sum_{k \in W} s'_{ik} \,. \tag{13}$$

The row i, for which the maximum is found is assumed to represent the topic of sentence being recognised. The calculated sum p_{sem} can be used as additional weight in providing speech recognition due to Bayes' theorem, however, it is not a probability function because $p_{sem} \in \Re^{+}$. 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_{sem} received from semantic model. The final measure can be obtained applying Bayes' theorem

$$p = p_{htk}^{w} p_{sem} \,. \tag{14}$$

7. Experimental Results

We experimented using different values of parameters n, α and w. Results for recognition based on audio model only are also included. LSA was used as a baseline to evaluate results of our method. We experimented with several different w for semantic model based on LSA and values in a range between 23 and 26 gave the best result presented in Tab. 3. The 45 utterances did not have hypotheses with correct sentences in entire

Table 3. Experimental results for pure HTK audio model, audio model with LSA and audio model with our semantic model

n	*α*	*w*	recognised sentences	%
LSA		25	41	0.36
HTK			33	0.29
3	1	50	48	0.42
3	2	50	46	0,40
3	3	50	46	0.40
7	1	50	35	0.31
7	3	50	45	0.39
7	5	50	46	0.40
5	1	20	44	0.39
5	2	20	55	0.48
5	3	20	60	0.53
5	4	20	59	0.52
5	5	20	59	0.52
3	2	20	61	0.53
3	1	20	50	0.44
7	6	20	59	0.52
7	5	20	61	0.53
7	4	20	59	0.52
8	4	20	57	0.50
8	5	20	61	0.53
8	6	20	60	0.53
9	1	20	28	0.25
9	3	20	49	0.43
9	5	20	57	0.50
9	6	20	61	0.53
9	7	20	59	0.52
11	5	20	54	0.47
11	7	20	60	0.53
11	8	20	60	0.53
11	9	20	58	0.51
9	6	10	58	0.51
9	6	15	60	0.53
9	6	17	60	0.53
9	6	18	61	0.53
9	6	19	61	0.53
9	6	20	61	0.53
9	6	22	59	0.52
9	6	25	58	0.51

10 best lists. This is why the maximal number of utterances which were possible to be recognised is 69.

The experiment shows that our semantic model is useful. Results might be so outstanding due to a small number of words in the corpus and using the same corpus for training and testing. The same corpus was used for both tasks because phoneme segmentation in the corpus is needed to use HTK. CORPORA is the only Polish corpus which provides it. However, the comparison of 53% correct recognitions for best configurations of our model with 36% for LSA and 29% for audio model only is impressive. The analysed results for different configurations shown that the choice of n, the length of list of topics related to an analysed topic is not as important as ratio between n and α which is a smoothing factor for weighting impact of related topics. The ratio n/α should be kept around 2/3 in order to provide the best results. The audio model importance weight w is also very crucial as the information from HTK model is very important and can be ignored if w has too small value.

It has to be mentioned that this was a preeliminary experiment. Our aim was to check if there is point in spending more time on reserach on this model. This is why we used little data and the same set for training and testing. We do not claim that the calculated model can be used in any practical task. One more reason is that it was trained on CORPORA which has no semantic connotations. On the other hand it has to be stressed that for Polish this model keeps some grammar information as well eventhough we called him a semantic one. In example we can expect words with morphology related to one gender in a given small set of words, which will be noted in the matrix. We found the results promising and now we work on a model describing Polish language properly using transciptions from Polish Parliment as a corpus.

Another way of proving usefulness of our semantic model is through calculating histograms p_{sem} of probabilities received from semantic model for hypotheses which are correct recognitions (Fig. 2) and histogram p_{semw} of probabilities received from semantic model for hypotheses which are wrong recognitions (Fig. 3). The ratio $p_{semc}/(p_{semc} + p_{semw})$ is presented in Fig. 4. It clearly shows a correlation between high probability from a semantic model and correctness of recognition.

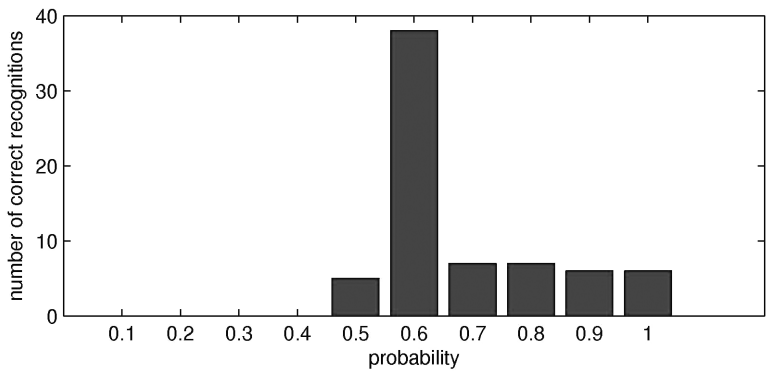

Figure 2. Histogram of probabilities received from semantic model for hypotheses which are correct recognitions.

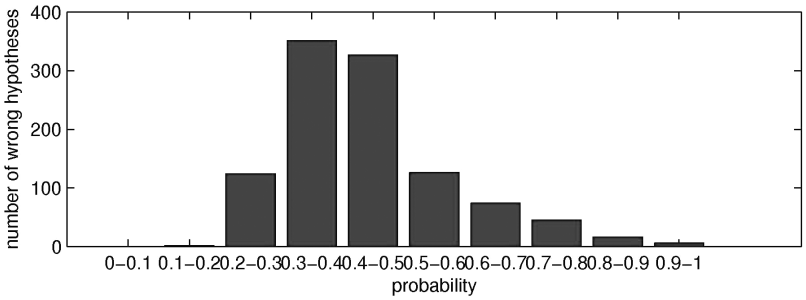

Figure 3. Histogram of probabilities received from semantic model for hypotheses which are wrong recognitions.

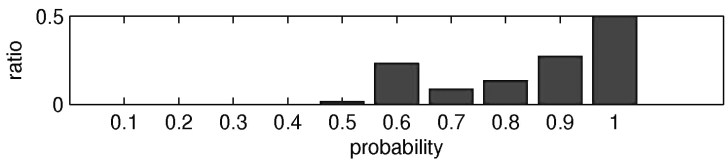

Figure 4. Ratio of correct recognitions to all of them for different probabilities received from semantic model.

8. Conclusions

The new method inspired by LSA was presented. The experimental results provide 17% higher correct recognition than LSA. The advantage of the method is that smoothing of information in a matrix representing word-topics relations is based on a limited number of most closely related topics for every topic rather than on all of them like in LSA. Applying the method on a level above the HTK model gave almost 2 times more correct recognitions than pure audio model. The method has too be trained and tested on larger and separate corpora to be used in practical applications.

BIBLIOGRAPHY

[1] Bellegarda, J. R. 1997. A latent semantic framework for large-span language modelling. [In:] *Proceedings of Eurospeech*, vol. 2 pp. 1451–1454.

[2] Agirre, E., Ansa, O., Martinez, D., Hovy, E. 2001. Enriching wordnet concepts with topic signatures [In:] *Proceedings of the SIGLEX Workshop on WordNet and Other Lexical resources: Applications, Extensions and Customizations*.

[3] Agirre,E. Alfonseca, E., de Lacalle, O. L., 2004. Approximating hierarchy-based similarity for wordnet nominal synsets using topic signatures [In:] *proceedings of the 2nd Global WordNet Conference*.

[4] Harary F., 1969. *Graph Theory*.

[5] Veronis J., 2004, Hyperlex: lexical cartography for information retrieval [In:] *Computer Speech and Language*.

[6] Agirre, E. Martinez, D., de Lacalle, O. L., Soroa, A., 2006. Two graph-based algorithms for state-of-the-art [In:] *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 585–593.

[7] Grocholewski, S., 1995. Założenia akustycznej bazy danych dla języka polskiego na nośniku cd rom. [In:] Mat. 1 KK: *Głosowa komunikacja człowiek-komputer*, pp. 177–180.

[8] Young, S., 1996. Large Vocabulary continuous speech recognition: a review. [In:] *IEEE Signal Processing Magazine*, vol. 13(5), pp. 45–57.

[9] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland P., 2005. *HTK Book*.

[10] Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*.