Semantically Based Search in a Social Speech Task

Fernando García, Emilio Sanchis, Ferran Pla

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camino de Vera s/n, 46020 Valencia, Spain
{fgarcia,esanchis,fpla}@dsic.upv.es

Abstract

In this work, we present an approach for semantically based search for similar segments of a speech task, i.e., the search for audio segments in a row audio repository that are semantically related to the audio segment given by a user. Our approach is based on the lexical representation of segments of words that are enriched by semantic relations. We have studied different distance measures and the lexical/semantic representation of the segments. We present experiments for a task of recorded dialogs between students talking about whatever they want, which is a semantically unbounded task. The results, which are encouraging, indicate the potential advantages of using this approach to address this problem.

Index Terms: audio search, semantic similarity, social speech.

1. Introduction

In the last few years, there has been increasing interest in many areas related to analysis of big repositories of documents and speech. This is the case for information retrieval, information extraction, classification, topic detection, etc. One interesting topic is the search and classification of documents based not only on lexical or acoustic similarity but also on the semantics of the documents.

Semantic text similarity is a widely studied topic in the natural language processing field. The main objective of this task is to measure the degree of semantic equivalence between two texts. Specific tasks have been recently defined to determine the semantic equivalence between two texts at the * SEM conference [1, 2]. Most approaches are based on distance measures that compute word similarities from WordNet [3] in different ways: by computing the overlaps of the words with the glosses of the synsets of the words considered by defining variants of the work of Lesk[4] or exploring the relationship between the synsets using the WordNet hierarchy and the WordNet relations to obtain the similarity of synsets based on the shortest path between two synsets. A well-known available resource that implements the most text similarity measures is the WordNet Similarity Package [5].

Also, in the field of speech processing, there are many applications for search and information extraction/classification [6]. This is the case for some of the tasks proposed at the MediaEval conference [7] such as Query by example search [8, 9, 10, 11, 12] or Similar Segments in Social Speech [13]. Another interesting task is the possibility of finding segments of speech based on some information provided to the system (e.g., in applications of topic detection). This information could be provided in terms of significant words or in a uninformed way, simply by means of a segment of speech that is relevant to this search. In other words, the goal is to find segments of speech that are semantically similar. An important characteristic is that this is an open domain task, which means that specific modelizations cannot be used for the domain. This was a new task in the last MediaEval conference, where some approaches were based on lexical distance similarity and other approaches used prosodic information [14, 15, 16, 17]. The corpus consists of conversations between students that were recorded at a university department. They talked about whatever they wanted, and the task was to return a list of jump-in points of regions semantically similar to it given a short audio segment of interest.

In this paper, we present an approach for finding similar segments of speech. This is an extension of the system we presented at the MediaEval 2013 workshop and achieved good results [13, 14]. The approach we present is mainly based on lexical similarity, but the comparison is enriched by the semantics of words. In other words, by using a priori linguistic knowledge (such as synonyms, hypernyms, hyponyms,..., extracted from WordNet) the segments are better characterized than when only words are used. This way we can achieve two goals. On the one hand, the coverage in the search is increased by the effect of augmenting the number of words that can be similar to the test segment words; on the other hand, the semantic ambiguity of words is reduced by the effect of including a kind of semantic representation. We have studied several distance definitions and several segment representations. Finally, we present the experiments conducted and a discussion of the obtained results.

2. Task Description and data set

The goal of the task is to find a list of jump-in points in an audio/video file that are semantically similar to a short audio/video region of interest. The corpus consisted of approximately five hours of dialogs among students with no limitations on the topics that they could talk about. The corpus was divided into a training set and a test set. The training set consisted of 20 dialogs (241 minutes in total) with the most common topics relating to classes and class assignments, interesting new technologies, career ambitions, games, and movies. The test set consisted of 6 dialogs (68 minutes in total). Even though the topics were less about classes and more about research, there are otherwise fairly similar.

The annotations are tagsets which indicate regions that are similar in some way. The average duration was 50 seconds in the training set, and after clarifying the instructions to the annotators the average duration of each segment was 31 in the test set. There were 198 tagsets over the training set, with a total of 1697 tagged regions, and 29 tagsets and 189 tagged regions for the test set.

Some of the topics in the corpus are: *#food, #travel, #cars-*

and-driving, #planning-class-schedules, #TV-shows, #lack-of-money, #family, #anecdotes, #problems, #short-term-future-plans, #advice, #gossip, and *#positive-things-about-classes*.

It should be noted that the topics present in the test set corpus are different than those in the training set. Therefore, a topic search based on models learned from the training corpus is not suitable.

3. System Description

Our approach has two phases (Figure 1). In the first phase, a representation (bag-of-words) that only contains the most relevant words in the corpora is found, and in the second phase similarity distances between these bag-of-words are calculated.

The corpus was divided into segments of words. For the human transcriptions, the segments were the sentences that were produced by the transcribers, and, for the ASR transcriptions the segments were pause-delimited regions. The segments include the queries and the potential results.

In the first phase, we received the segments of words provided by the manual transcription and we received the output of the Automatic Speech Recognition (ASR). We did a part-of-speech (POS) tagging [18], in order to disambiguate some words and to set up the segments of words for the next step. Then we applied morphological processing to get the base forms of the words since inflections (number, tense, etc.) were not generally semantically relevant. Afterwards the segments of words were filtered, using a large list of approximately 500 stopwords, including standard stopwords and some words that are typical of spontaneous speech. Finally, we obtained the bag-of-words representation of the segments of words.

In the second phase we compared the bag-of-words of each query and the bag-of-words of each possible result to obtain a score that was associated to the starting point of that segment. We tried several methods to compute similarity [14] and we found that measures based on the number of words that are common to both bag-of-words worked well. Specifically these included the $dot\ product$, the $cosine\ similarity$, and the $multiset\ intersection$ when representing the segments as vectors of word frequencies:

$$dot\ product = \overrightarrow{q} \cdot \overrightarrow{s} = \sum_{i=1}^{|V|} q_i s_i \tag{1}$$

$$cosine\ similarity = \frac{\sum_{i=1}^{|V|} q_i s_i}{\sum_{i=1}^{|V|} (q_i)^2 \cdot \sum_{i=1}^{|V|} (s_i)^2} \tag{2}$$

$$multiset\ intersection = \sum_{i=1}^{|V|} min(q_i, s_i) \tag{3}$$

where V is the vocabulary and $\overrightarrow{q}$, $\overrightarrow{s}$ are the word frequency vectors representing the query and the segment, respectively.

These measures can be expected to work well in terms of precision since they rank the segments that share many words highly.

In order to improve the coverage, we augmented it with measures that take into account lexical and semantic generalizations. These were based on information in WordNet using the software package WordNet Similarity [5] that gives a measure of the semantic similarity and relatedness between any pair of words. For these experiments, the measures that we used were the following: two similarity measures based on path lengths between synsets: (`lch`, `wup`), another two measures based on information content (`lin`, `jcn`) and the `lesk` measure.

- The `lch` and `wup` measures are defined taking into account the is_a relation from WordNet to determine the shortest path between two synsets. This hierarchical relation is only established between words that belong to the same POS, i.e., noun-noun relations, and verb-verb relations. Despite the fact that other relations exist in WordNet, only these hierarchical relation were applied in these measures.
- The `lin` and `jcn` measures use external resources (the sense-tagged $Semcor$ corpus) to obtain the frequency of the synsets in the corpus. This information is useful for augmenting the information content of the least common subsumer (LCS) measure. LCS found the most specific synset that is an ancestor of two synsets in an is_a relation.
- The `lesk` measure uses the glosses associated to each synset of WordNet. It computes the overlaps between the glosses of the two synsets to compute the relatedness between them.

We combined these measures with the previous scores by linear interpolation as follows:

$$\kappa \cdot \begin{bmatrix} dot\ product \\ cosine\ similarity \\ multiset\ intersection \end{bmatrix} (q,s) + (1-\kappa) \cdot \begin{bmatrix} jcn \\ lch \\ lesk \\ lin \\ wup \end{bmatrix} (q,s) \tag{4}$$

These combined measures improved the results due to inclusion of the semantic relations between words extracted from WordNet database.

4. Evaluation of the results

For evaluation purposes, some specific measures were defined in [13]. These specific measures are based on the characteristics of the task. In other words, it is assumed that the system will provide a set of jump-in points and the user must browse around the points suggested because the segment boundaries cannot be strongly delimited. In order to adapt standard measures to this scenario some metrics were proposed. A jump-in point was considered to be a hit if it was not more than 5 seconds before the start of a relevant region and no more than 3 seconds before the end. A measure of how users can use the suggested jump-in points is the so-called Searcher Utility Ratio (SUR), which is a variant of [19, 20] when the data is not pre-segmented. It is a measure of the number of seconds that a searcher needs to explore around the suggested jump-in points. The numerator of the SUR is the estimated value by the searcher, and it is the number of seconds of relevant audio/video that can be found by using the suggested jump-in points; the denominator is the estimated cost, also measured in seconds. There are three cases:

1. If the suggested jump-in point does not correspond to any ground-truth region (a false-positive error), then the cost is 8 seconds, which is the estimated time that a searcher needs to determine that it is a false alarm.
2. If the suggested jump-in point is no more than 5 seconds before the actual region start point, the cost is the time from that jump-in point to the end of the actual region.

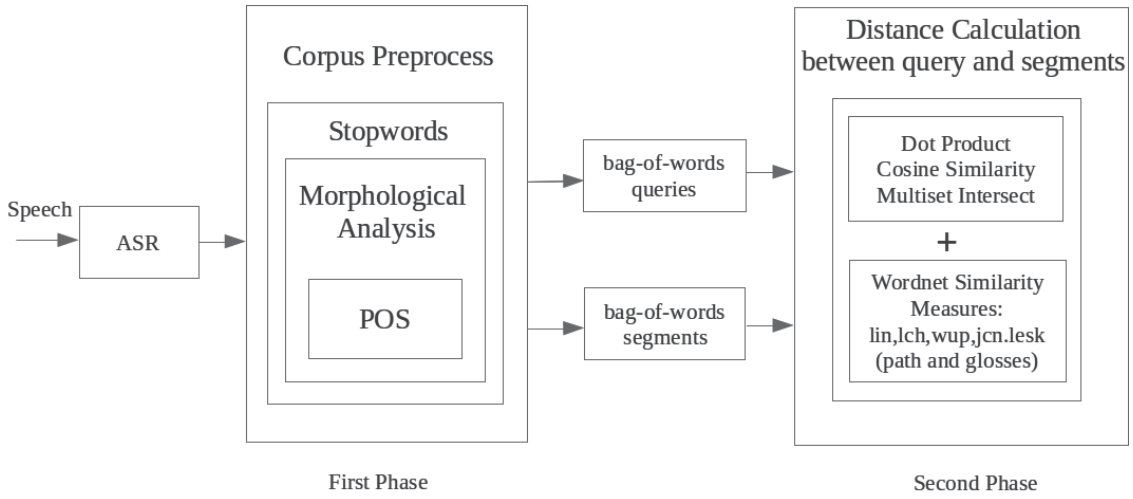

Figure 1: Scheme of our approach.

This reflects the time spent to scan forward to the start of the relevant content and the time spent to listen to it.

3. If the suggested jump-in point is within the region, then the benefit is the remaining duration of the region and the cost is the same.

Recall is the fraction of obtainable content actually found, where the obtainable content is the total content in the other regions in the tagset, up to a maximum of two minutes

The overall measure is the weighted F-measure, F, with the Searcher Utility Ratio weighted higher than the Recall:

$$F = \frac{10 \cdot nsur \cdot nr}{nsur + 9 \cdot nr} \quad (5)$$

where $nsur$ is the normalized Searcher Utility Ratio and nr is the normalized Recall.

5. Experimental results

The training set was 20 dialogs (241 minutes in total). The test set was 6 dialogs (68 minutes in total). The average durations of the segments of words were 50 seconds in the training set and 31 in the test set.

In order to evaluate the output of the system, some measures were defined:

```
fa = false alarms
hits = hits
early = number of exact or early jump-in points
late = number of late jump-in points
np = naive Precision percentage = hits/(hits+fa)
rseu = raw Searcher Utility Ratio
nsur = normalized Searcher Utility Ratio
nr = normalized Recall
F = weighted F-measure
```

where the weighted F-measure is the the overall measure.

Table 1 shows the results for the different lexical measures for both the human transcriptions and the output of the ASR training corpora. As the table indicates, the naive Precision percentage (number of hits / number of responses) was the same for the $dot\ product$ and the $cosine\ similarity$ and for both corpora. The different scores for the normalized Searcher Utility Ratio (nsur) and the normalized Recall (nr) in the first two measures made the score of F slightly higher for the second measure in the human transcriptions. For the output of the ASR the different scores for nsur and nr in the first two measures causes that the score of F was slightly higher for the first measure. Finally in the case of the $multiset\ intersection$ measure the scores were the worst for both corpora.

Table 1: *The results for different lexical distances for the training corpora.*

Human Transcriptions				
score	np	nsur	nr	F
dot product	12.8%	1.26	0.58	1.13
cosine similarity	12.8%	1.30	0.65	1.16
multiset intersect	12.7%	1.27	0.58	1.13
ASR				
score	np	nsur	nr	F
dot product	8.4%	0.93	0.64	0.89
cosine similarity	8.4%	0.96	0.48	0.87
multiset intersect	8.2%	0.92	0.44	0.83

Table 2: *The results for the lexical distances combined with the WordNet measures.*

human transcriptions				
score	np	nsur	nr	F
dot product+wn	13.4%	1.286	0.600	1.15
cosine similarity+wn	13.1%	1.319	0.610	1.18
multiset intersect+wn	12.9%	1.296	0.744	1.21
ASR				
score	np	nsur	nr	F
dot product+wn	8.2%	0.934	0.675	0.90
cosine similarity+wn	9.3%	1.065	0.456	0.94
multiset intersect+wn	8.0%	0.887	0.781	0.88

Table 2 shows the results for the human transcriptions and the ASR training corpus using the different lexical distance measures combined with the five WordNet semantic similarity measures (wn). As the table indicates, the results outperformed the results shown in Table 1 for each of the measures proposed;

Table 4: *The results for the test corpora.*

Lexical distances for the human transcriptions										
	fa	hits	early	late	np	rr	rsur	nsur	nr	F
dot product	70	14	6	8	16.7%	20%	0.41	1.41	0.71	1.28
cosine similatity	69	15	8	7	17.9%	25%	0.47	1.63	0.91	1.51
multiset intersection	68	13	5	8	16.0%	19%	0.40	1.38	0.67	1.25
Lexical distances combined with WordNet measures for the human transcriptions										
	fa	hits	early	late	np	rr	rsur	nsur	nr	F
dot product+wn	65	18	8	10	21.7%	28%	0.46	1.59	1.01	1.51
cosine similatity+wn	66	17	10	7	20.5%	25%	0.43	1.49	0.91	1.40
multiset intersection+wn	84	17	8	9	16.8%	21%	0.38	1.27	0.75	1.19
Lexical distances for the ASR output										
	fa	hits	early	late	np	rr	rsur	nsur	nr	F
dot product	119	14	5	9	10.5%	25%	0.34	1.18	0.92	1.15
cosine similatity	74	6	4	2	7.5%	11%	0.28	0.92	0.39	0.81
multiset intersection	156	16	5	11	9.0%	24%	0.28	0.95	0.88	0.95
Lexical distances combined with WordNet measures for the ASR output										
	fa	hits	early	late	np	rr	rsur	nsur	nr	F
dot product+wn	88	14	7	7	13.7%	27%	0.43	1.47	1.00	1.40
cosine similatity+wn	134	12	7	5	8.2%	20%	0.27	0.94	0.74	0.92
multiset intersection+wn	127	16	5	11	11.2%	27%	0.34	1.17	0.97	1.14

Table 3: *The F results for the different lexical measures combined with each of the WordNet measures.*

human transcriptions					
score	jcn	lch	lesk	lin	wup
dot product	1.18	1.13	1.19	1.13	1.16
cosine similarity	1.19	1.20	1.20	1.21	1.18
multiset intersection	1.25	1.22	1.18	1.20	1.22
ASR					
score	jcn	lch	lesk	lin	wup
dot product	0.90	0.89	0.89	0.87	0.91
cosine similarity	0.88	0.91	0.92	0.90	0.94
multiset intersection	0.88	0.90	0.90	0.87	0.87

therefore, the results were better when WordNet was used to calculate the distances.

To evaluate the influence of the different semantic measures calculated in WordNet, we did some experiments using the linear combination of the lexical measure with each of the WordNet similarity measures calculated in this paper, the results are shown in Table 3 in terms of the weighted F-measure.

For the test set, the results for both the human transcriptions and the ASR output are shown in Table 4. In both cases (human, ASR), the best results were obtained with the combination of the lexical measure $dot\ product$ and the average of the five WordNet measures $dot\ product + wn$ with a value of $\kappa = 0.7$, four responses were provided for the human test set and seven responses were provided for the ASR set for each of the 21 test set queries. It should be noted that for the results of the $cosine\ distance$ and the $multiset\ intersection$ distances the F value was the worst; nevertheless the number of hits and consequently the $naivePrecision$ (np) increased. Therefore, we observe that the use of measures calculated in WordNet help lexical distances to achieve better results especially when there are recognition errors.

In the case of the results for ASR output, the $dot\ product + wn$ lexical distance obtained better results than the other two measures. This means that the measure that takes into account the repetition of words works better than the other two smoothed measures, which calculates distances between words without repeating or which calculates a minimum set as in the case of multiset intersection.

The results for the ASR output were not too different from the results obtained using the human transcriptions. This may be because our similarity measure is strongly based on relevant words which can be better recognized than many short stopwords which are removed by our process.

6. Conclusions

In this paper, we have presented an approach for a task of finding semantically similar segments in an audio/video repository. We have addressed the problem from the point of view of a semantical/lexical classification after the ASR process. Although the lexical-based approach obtains reasonable results (given the difficulty of the task), the semantic characterization obtained from WordNet can improve the results. This may be because the coverage of the model is increased when a word is not just presented by itself, but it is also represented with synonyms or by words semantically related by hypernyms, hyponyms and so on relations. Also, these informations allow semantically ambiguous words to match better with words in the same semantic context.

For future work we plan to study other different measures. It may also be interesting to explore other representations of the segments that are different from a bag-of-words in a more structured way where we can represent relations between words and weights in order to represent the relevance of words in segments.

7. Acknowledgements

We want to thank Nigel G. Ward and Steven D. Werner for their suggestions and comments during this work. This work is funded by the Spanish Government under the contracts TIN2011-28169-C05-01 and TIN2012-38603-C02-01.

8. References

[1] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (* SEM 2012)*, 2012.

[2] E. Agirre, D. Cer, M. Diab, A. Gonzalez-agirre, and W. Guo, "sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity," in *In *SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics*, 2013.

[3] G. A. Miller, "WordNet: A Lexical Database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. [Online]. Available: http://doi.acm.org/10.1145/219717.219748

[4] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proceedings of the 5th Annual International Conference on Systems Documentation*, ser. SIGDOC '86. New York, NY, USA: ACM, 1986, pp. 24–26. [Online]. Available: http://doi.acm.org/10.1145/318723.318728

[5] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Measuring the Relatedness of Concepts," in *Proc. of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, 2004, pp. 1024–1025.

[6] M. Larson and G. J. F. Jones, "Spoken content retrieval: A survey of techniques and technologies." *Foundations and Trends in Information Retrieval*, vol. 5, no. 4-5, pp. 235–422, 2012.

[7] MediaEval Benchmarking Initiative for Multimedia Evaluation., http://www.multimediaeval.org/medieval2013.

[8] L.-F. Hurtado, M. Calvo, J. A. Gómez, F. García, and E. Sanchis, "A Phonetic-Based Approach to Query-by-Example Spoken Term Detection," in *The 18th Iberoamerican Congress on Pattern Recognition (CIARP)*, Havana, Cuba, November 20-23 2013.

[9] J. A. Gómez, L.-F. Hurtado, M. Calvo, and E. Sanchis, "ELiRF at MediaEval 2013: Spoken Web Search Task," in *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[10] F. Metze, E. Barnard, M. Davel, C. Van Heerden, X. Anguera, G. Gravier, N. Rajput *et al.*, "The spoken web search task," in *Working Notes Proceedings of the MediaEval 2012 Workshop*, 2012.

[11] E. Sanchis, L.-F. Hurtado, J. A. Gómez, M. Calvo, and R. Fabra, "The ELiRF Query-by-Example STD systems for the Albayzin 2012 Search on Speech Evaluation," in *Procedings of Iberspeech 2012*, Madrid, Spain, November 21-23 2012.

[12] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational*. Citeseer, 2007, pp. 51–55.

[13] N. G. Ward, S. D. Werner, D. G. Novick, E. E. Shriberg, C. Oertel, L.-P. Morency, and T. Kawahara, "The Similar Segments in Social Speech Task," in *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[14] F. García, E. Sanchis, M. Calvo, F. Pla, and L.-F. Hurtado, "ELiRF at MediaEval 2013: Similar Segments in Social Speech Task," in *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[15] P. Galuščáková and P. Pecina, "CUNI at MediaEval 2013 Similar Segments in Social Speech Task," in *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[16] S. D. Werner and N. G. Ward, "Evaluating Prosody-Based Similarity Models for Information Retrieval," in *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[17] G.-A. Levow, "UWCL at MediaEval 2013: Similar Segments in Social Speech Task," in *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[18] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," in *In Proceedings of HLT-NAACL*, 2003, pp. 252–259.

[19] B. Liu and D. W. Oard, "One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06. New York, NY, USA: ACM, 2006, pp. 673–674. [Online]. Available: http://doi.acm.org/10.1145/1148170.1148311

[20] M. Eskevich, R. Aly, R. Ordelman, S. Chen, and G. J. Jones, "The Search and Hyperlinking Task at MediaEval 2013," in *The MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.