ELiRF at MediaEval 2013: Similar Segments in Social Speech Task

Fernando García, Emilio Sanchis, Marcos Calvo, Ferran Pla, Lluís-F. Hurtado
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46020, València, Spain
{fgarcia,esanchis,mcalvo,fpla,lhurtado}@dsic.upv.es

ABSTRACT

This paper describes the Natural Language Engineering and Pattern Recognition group (ELiRF) approaches and results towards the Similar Segments of Social Speech Task of MediaEval 2013. The task involves finding segments similar to a query segment in a multimedia collection of informal, unstructured dialogs among members of a small community. Our approach has two phases. In a first phase a preprocess of the sentences is performed based on the morphology and semantics of the words. In a second phase, a searching process based on different distance measures is carried out. This has been done taking the correctly transcribed sentences and the output of an Automatic Speech Recognizer.

1. INTRODUCTION

The Similar Segments of Social Speech Task of MediaEval 2013 [3] involves searching in social multimedia. The corpus consists of conversations between students in a university department. This task is the first exploration of social search in multimedia, and the first social spoken dialog retrieval task not assuming term-based search.

The corpus given by the organization consisted of a 5-hour collection of dyadic English-language conversations (4 for training and 1 for test), each 5-10 minutes in length, by members of a semi-cohesive group.

The input to the systems is a 1-10 second audio/video region of interest, and the desired output is an ordered list of regions similar to it, matching as closely as possible the judgments of human searchers.

2. SYSTEM DESCRIPTION

Our approach consists of two phases: a first one to obtain an accurate representation of the query segments and the test segments, and a second one to compare the query representation with the succesive test segments.

Figure 1 represents the architecture of the system. Our lexical/semantic modeling and distance calculations are based on words, thus, we start from the output of a previous ASR process that provides a single sentence. It should be noted that for this kind of tasks the quality of the ASR is very important, and it must be robust enough to give reasonable results for open vocabulary tasks. The **first phase** of our system is the same for the query segments and for the test segments. Its goal is to find the words that better represents the meaning of the segment.To do this we have performed the following processes: POS-tagging, morphological processing, stopwords filtering.

- POS-tagging: We associate a POS tag with each word. To do this we have used the Stanford POS-tagger [2]. This information will be useful to detect relevant words (nouns, verbs,...) and to help the following morphological process.
- Morphological process: Given that some morphological inflections are not relevant for the meaning (gender, singular, plural, verb inflections,..) it is convenient to obtain base forms of the words. To achieve this, once the POS-tagging is performed, the WordNet tool "Morphy"[1] that provides the base form of the words (considering the POS tag) was used.
- Stopwords filtering: In order to remove the unrelevant words for the representation of the meaning of the sentences, a list of stopwords was used. It is a classical stopwords list enriched with some words that can typically appear in spontaneous speech.

Then each segment is represented by the bag-of-words obtained from it. We consider that for this task it can be necessary to have a generalization mechanism that permits the detection of similar segments even when they have no words in common. This is the case of the use of synonyms when talking about the same things, or different specifical aspects of a more general topic. In our system, instead of including this generalization in the sentence representation, we have used a comparison mechanism in the second phase that takes into account this kind of lexical/semantic generalization.

The **second phase**: Once the bag-of-words for the query and for the test segments are obtained, the second phase compares these bag-of-words given and provides as a result the beginning time of the segments that can be considered similar to the query (jump-in points). We have used several similarity measures. The simplest one is just to find the number of common words in both bag-of-words (w2w).

$$\max_{\forall segment} |query \cap segment|$$

This measure is expected to work well in terms of Precision, as it detects the segments that share a lot of words,

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

[1] http://wordnet.princeton.edu/

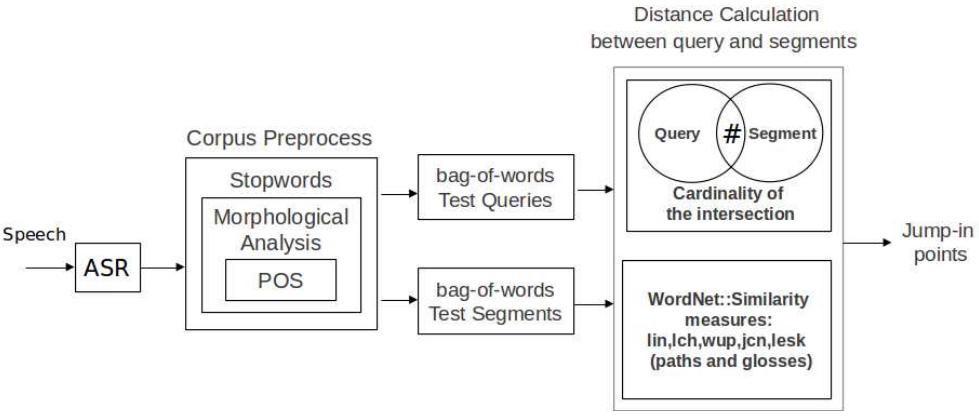
Distance Calculation
between query and segments
Corpus Preprocess
Stopwords
Morphological Analysis
POS
Speech
ASR
bag-of-words Test Queries
bag-of-words Test Segments
Query
#
Segment
Cardinality of the intersection
WordNet::Similarity measures: lin,lch,wup,jcn,lesk (paths and glosses)
Jump-in points

Figure 1: Scheme of our approach.

Table 1: Results for Human and ASR test corpora.

	FA	hits	early	aeo	late	alo	np	rr	rsur	nsur	nr	F
Human w2w	70	14	6	1.9	8	8.1	17%	20%	0.408	1.408	0.710	1.28
Human w2w_wn	73	11	3	1.4	8	11.2	13%	17%	0.374	1.289	0.632	1.17
ASR w2w	119	14	5	2.7	9	14.0	11%	25%	0.342	1.178	0.924	1.15
ASR w2w_wn	95	9	3	2.0	6	10.6	9%	15%	0.276	0.950	0.528	0.88

but it can not generalize to include the diferents ways of to talk about similar topics.

In order to have more coverage, we have explored some measures that take into account lexical and semantic generalizations. These measures are based in the information contained in WordNet. We have used the software package WordNet::Similarity [1] that permits to measure the semantic similarity and relatedness between a pair of words.

For this experiments the measures we used were: two similarity measures based on path lengths between concepts: (`lch`, `wup`), other two based on information content (`lin`, `jcn`) and the `lesk` measure, that uses the text of the dictionary gloss as a unique representation for the underlying concept.

Considering these measures and the previous one we have defined a new measure that is a linear combination of them (w2w_wn):

$$\max_{\forall segment} \lambda \cdot |query \cap segment| +$$

$$+(1-\lambda) \cdot (lin + lch + wup + jcn + lesk)$$

3. EXPERIMENTS

The test set consisted in a set of 6 dialogs (68 minutes) and a set of 21 regions of interest, or seeds. For each seed, the system should return a list of jump-in points representing the inferred similar-regions.

The task data set includes two transcriptions of the corpus: a manual transcription (Human) and a transcription obtained by an ASR (ASR). Both of them consist of a sequence of segments of words and the beginning and ending time associated with them.

In order to evaluate the output of the systems, the official metrics for the task are [3]: `FA` = false alarms, `hits` = hits, `early` = number of exact or early hits, `aeo` = average early offset, `late` = number of late hits, `alo` = average late offset, `rr` = raw recall, `rseu` = raw searcher utility ratio, `nsur` = normalized searcher utility ratio, `nr` = normalized recall, `F` = F-measure.

In Table 1 we show our results for both the human and the automatically transcribed test set corpora. In the case of the w2w_wn experiments, the value of λ that provides the best results is 0.7. As Table 1 shows, the results are worse when WordNet is used to calculate the distances. This can be due to the over-generalization generated by the semantic similarities found in WordNet. Furthermore, the results with the ASR output are not too far from the results using the human transcription. This could happen 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.

4. ACKNOWLEDGMENTS

Work funded by the Spanish Government and the E.U. under the contracts TIN2011-28169-C05 and TIN2012-38603-C02, and FPU Grant AP2010-4193.

5. REFERENCES

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

[2] 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*, pages 252–259, 2003.

[3] 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.