SAIL-GRS: Grammar Induction for Spoken Dialogue Systems using CF-IRF Rule Similarity

Kalliopi Zervanou, Nikolaos Malandrakis and Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory (SAIL),
University of Southern California, Los Angeles, CA 90089, USA
`kzervanou@gmail.com`, `malandra@usc.edu` , `shri@sipi.usc.edu`

Abstract

The SAIL-GRS system is based on a widely used approach originating from information retrieval and document indexing, the $TF\text{-}IDF$ measure. In this implementation for spoken dialogue system grammar induction, rule constituent frequency (CF) and inverse rule frequency (IRF) measures are used for estimating lexical and semantic similarity of candidate grammar rules to a seed set of rule pattern instances. The performance of the system is evaluated for the English language in three different domains, travel, tourism and finance and in the travel domain, for Greek. The simplicity of our approach makes it quite easy and fast to implement irrespective of language and domain. The results show that the SAIL-GRS system performs quite well in all three domains and in both languages.

1 Introduction

Spoken dialogue systems typically rely on grammars which define the semantic frames and respective fillers in dialogue scenarios (Chen et al., 2013). Such systems are tailored for specific domains for which the respective grammars are mostly manually developed (Ward, 1990; Seneff, 1992). In order to address this issue, numerous current approaches attempt to infer these grammar rules automatically (Pargellis et al., 2001; Meng and Siu, 2002; Yoshino et al., 2011; Chen et al., 2013).

The acquisition of grammar rules for spoken language systems is defined as a task comprising of two subtasks (Meng and Siu, 2002; Iosif and Potamianos, 2007), the acquisition of:

(i) ***Low-level rules*** These are rules defining domain-specific entities, such as names of locations, hotels, airports, e.g. `CountryName`: “USA”, `Date`: “July 15th, 2014”, `CardType`: “VISA” and other common domain multi-word expressions, e.g. `DoYouKnowQ`: “do you know”.

(ii) ***High-level rules*** These are larger, frame-like rule patterns which contain as semantic slot fillers multi-word entities identified by low-level rules. For example: `DirectionsQ`: “`<DoYouKnowQ>` `<where>` the `<MuseumName>` is located”, `ExpressionCardProblem`: “my `<CardType>` has expired”.

The shared task of *Grammar Induction for Spoken Dialogue Systems*, where our system participated, focused on the induction of high-level grammar rules and in particular on the identification and semantic classification of new rule patterns based on their semantic similarity to known rule instances.

Within this research framework, the work described in this paper proposes a methodology for estimating rule semantic similarity using a variation of the well-known measure of $TF\text{-}IDF$ as rule constituent frequency vs. inverse rule frequency, henceforth $CF\text{-}IRF$.

In the remainder of this paper, we start in Section 2 by a detailed description of our system. Subsequently, in Section 3, we present the datasets used and the evaluation process, and in Section 4 we discuss our results. We conclude in Section 5 with a summary of our observations and directions for future work.

2 System Description

The SAIL-GRS system is based on a widely used approach in information retrieval and document indexing, the $TF\text{-}IDF$ measure. $TF\text{-}IDF$ is an approach that has found numerous applications in information management applications, such as document keyword extraction, (e.g., Dillon and Gray (1983)), document clustering, summarisation, (e.g., Gong and Liu (2001)), event clustering, (e.g., De Smet and Moens (2013)). In dialogue systems, $TF\text{-}IDF$ has been used, among other applications, for discovering local coherence (Gandhe and Traum, 2007) and for acquiring predicate-argument rule fragments in an open domain, information extraction-based spoken dialogue system (Yoshino et al., 2011). In their approach, Yoshino et al. (2011) use the $TF\text{-}IDF$ measure to determine the importance of a given word for a given domain or topic, so as to select the most salient predicate-argument structure rule patterns from their corpus.

In our implementation for spoken dialogue system grammar induction, rule constituent frequency (CF) and inverse rule frequency (IRF) measures are used for estimating lexical and semantic similarity of candidate grammar rules to a seed set of rule pattern instances. As illustrated in Table 1, the SAIL-GRS algorithm has two main steps, the training stage and the rule induction stage.

Input: known rule pattern instances **Output:** new candidate rule patterns
Training stage:
1. Known rule instance parsing 2. Rule constituent extraction (uni-/bigrams) 3. Rule constituent frequency count (CF) 4. Inverse rule frequency count (IRF) 5. $CF\text{-}IRF$ rule instance vector creation
Rule induction stage:
1. Unknown text fragment parsing 2. Unigram & bigram extraction 3. Uni-/bigram $CF\text{-}IRF$ value lookup 4. Creation of $CF\text{-}IRF$ vector for unknown text fragment 5. Estimation of cosine similarity of unknown fragment to rule instances 6. New candidate rule selection & rule semantic category classification using maximum cosine similarity

Table 1: The SAIL-GRS system algorithm

In the first, the *Training stage*, known rule instances are parsed and, for each rule semantic category, the respective high-level rule pattern instances are acquired. These patterns are subsequently split into unigram and bigram constituents and the respective constituent frequencies and inverse rule frequencies are estimated. Finally, for each rule category, a vector representation is created for the respective rule pattern instance, based on the $CF\text{-}IRF$ value of its unigram and bigram constituents.

In the second step, the *Rule induction stage*, the unknown text fragments are parsed and split into unigrams and bigrams. Subsequently, we lookup the known rule instance unigram and bigram representations for potential lexical matches to these new unigrams and bigrams. If these are found, then the new n-grams acquire the respective $CF\text{-}IRF$ values found in the training instances and the respective $CF\text{-}IRF$ vector for the unknown text fragments is created. Finally, we estimate the cosine similarity of this unknown text vector to each known rule vector. The unknown text fragments that are most similar to a given rule category are selected as candidate rule patterns and are classified in the known rule semantic category. An unknown text fragment that is selected as candidate rule pattern is assigned only to one, the most similar, rule category.

3 Experimental Setup

The overall objective in spoken dialogue system grammar induction is the fast and efficient development and portability of grammar resources. In the *Grammar Induction for Spoken Dialogue Systems* task, this challenge was addressed by providing datasets in three different domains, travel, tourism and finance, and by attempting to cover more than one language for the travel domain, namely English and Greek.

As illustrated in Table 2, the travel domain data for the two languages are comparable, with 32 and 35 number of known rule categories, for English and Greek, comprising of 982 and 956 high-level rule pattern instances respectively. The smallest dataset is the finance dataset, with 9 rule categories and 136 rule pattern instances, while the tourism dataset has a relatively low number of rule categories comprising of the highest number of rule pattern instances. Interestingly, as indicated in the column depicting the percent of unknown n-grams in the test-set, i.e. the unigrams and the bigrams without a $CF\text{-}IRF$ value in the training data, the tourism domain test-set appears also to be the one with the greatest overlap with the training data, with a mere 0.72% and 4.84% of unknown unigrams and bigrams respectively.

For the evaluation, the system performance is estimated in terms of precision (P), recall (R) and $F\text{-}score$ measures, for the correct classification of an unknown text fragment to a given rule category cluster of pattern instances. In addition to these measures, the weighted average of the per rule scores is computed as follows:

$$P_w = \frac{\sum_{i=1}^{N-1} P_i c_i}{\sum_{i=1}^{N-1} c_i}, \quad R_w = \frac{\sum_{i=1}^{N-1} R_i n_i}{\sum_{i=1}^{N-1} n_i} \quad (1)$$

$$F_w = \frac{2 \cdot P_w \cdot R_w}{P_w + R_w} \quad (2)$$

where $N-1$ is the total number of rule categories, P_i and R_i are the per rule i scores for precision and recall, c_i the unknown patterns correctly assigned to rule i, and n_i the total number of correct rule instance patterns for rule i indicated in the ground truth data.

4 Results

The results of the SAIL-GRS system outperform the Baseline in all dataset categories, except the Tourism domain, as illustrated in Table 3. In this domain, both systems present the highest scores compared to the other domains. The high results in the travel domain are probably due to the high data overlap between the train and the test data, as discussed in the previous section and illustrated in Table 2. However, this domain was also the one with the highest average number of rule instances per rule category, compared to the other domains, thus presenting an additional challenge in the correct classification of unknown rule fragments.

We observe that the overall higher F measures of the SAIL-GRS system in the travel and finance domains are due to higher precision scores, whereas Baseline system displays higher recall but lower precision scores and lower F-measure in these domains.

The overall lowest scores for both systems are reached in the Travel domain for Greek, which is also the dataset with the lowest overlap with the training data. However, the performance of the SAIL-GRS system does not deteriorate to the same extent as the Baseline, the precision of which falls to a mere 0.16-0.17, compared to 0.49-0.46 for the SAIL-GRS system.

5 Conclusion

In this work, we have presented the SAIL-GRS system used for the *Grammar Induction for Spoken Dialogue Systems* task. Our approach uses a fairly simple, language independent method for measuring lexical and semantic similarity of rule pattern instances. Our rule constituent frequency vs. inverse rule frequency measure, $CF\text{-}IRF$ is a modification the $TF\text{-}IDF$ measure for estimating rule similarity in the induction process of new rule instances.

The performance of our system in rule induction and rule pattern semantic classification was tested in three different domains, travel, tourism and finance in four datasets, three for English and an additional dataset for the travel domain in Greek. SAIL-GRS outperforms the Baseline in all datasets, except the travel domain for English. Moreover, our results showed that our system achieved an overall better score in precision and respective F-measure, in the travel and finance domains, even when applied to a language other than English. Finally, in cases of a larger percentage of unknown data in the test set, as in the Greek travel dataset, the smooth degradation of SAIL-GRS results compared to the Baseline indicates the robustness of our method.

A limitation of our system in its current version lies in the requirement for absolute lexical match with unknown rule unigrams and bigrams. Future extensions of the system could include rule constituent expansion using synonyms, variants or semantically or lexically similar words, so as to improve recall and the overall F-measure performance.

References

Yun-Nung Chen, William Yang Wang, and Alexander I. Rudnicky. 2013. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 120–125.

Wim De Smet and Marie-Francine Moens. 2013. Representations for multi-document event clustering. *Data Mining and Knowledge Discovery*, 26(3):533–558.

Martin Dillon and Ann S. Gray. 1983. FASIT: A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, 34(2):99–108.

Domain	High-Level Rule Categories #	Rule Patterns #		Test-set: Unknown n-grams %	
		Training-set	Test-set	Unigrams	Bigrams
Travel EN	32	982	284	5.13%	20.71%
Travel GR	35	956	324	17.26%	33.09%
Tourism EN	24	1004	285	0.72%	4.84%
Finance EN	9	136	37	12.35%	36.74%

Table 2: Characteristics of training and test datasets

Domain	SAIL-GRS						Baseline					
	P	P_w	R	R_w	F	F_w	P	P_w	R	R_w	F	F_w
Travel EN	0.57	0.54	0.66	0.62	0.61	0.58	0.38	0.40	0.67	0.69	0.48	0.51
Travel GR	0.49	0.46	0.62	0.51	0.55	0.49	0.16	0.17	0.73	0.65	0.26	0.26
Tourism EN	0.75	0.75	0.90	0.90	0.82	0.82	0.82	0.80	0.94	0.94	0.87	0.87
Finance EN	0.67	0.78	0.62	0.78	0.65	0.78	0.40	0.48	0.63	0.78	0.49	0.60

Table 3: Evaluation results for SAIL-GRS system compared to the baseline in all four datasets in terms of per rule Precision P, Recall R, and F-score F. In the grey column, P_w, R_w, and F_w stand for the weighted average of the per rule precision, recall and F-score respectively, as defined in Equ. 1 and 2.

Sudeep Gandhe and David Traum. 2007. First steps towards dialogue modelling from an un-annotated human-human corpus. In *Proceedings of the Fifth IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 22–27.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 19–25, New York, NY, USA. ACM.

Elias Iosif and Alexandros Potamianos. 2007. A soft-clustering algorithm for automatic induction of semantic classes. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association*, pages 1609–1612. ISCA.

Helen M. Meng and Kai-Chung Siu. 2002. Semi-automatic acquisition of semantic structures for understanding domain-specific natural language queries. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):172–181.

Andrew N. Pargellis, Eric Fosler-Lussier, Alexandros Potamianos, and Chin-Hui Lee. 2001. Metrics for measuring domain independence of semantic classes. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 447–450. ISCA.

Stephanie Seneff. 1992. TINA: A natural language system for spoken language applications. *Computational Linguistics*, 18(1):61–86, March.

Wayne Ward. 1990. The CMU air travel information service: Understanding spontaneous speech. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania*, pages 127–129.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. 2011. Spoken dialogue system based on information extraction using similarity of predicate argument structures. In *Proceedings of the SIGDIAL 2011 Conference*, pages 59–66. Association for Computational Linguistics.