tucSage: Grammar Rule Induction for Spoken Dialogue Systems via Probabilistic Candidate Selection

Arodami Chorianopoulou[†], **Georgia Athanasopoulou**[†], **Elias Iosif**[‡ †], **Ioannis Klasinas**[†], **Alexandros Potamianos**[*]
[†] School of ECE, Technical University of Crete, Chania 73100, Greece
[*] School of ECE, National Technical University of Athens, Zografou 15780, Greece
[‡] "Athena" Research Center, Marousi 15125, Greece
{achorianopoulou,gathanasopoulou,iklasinas}@isc.tuc.gr
iosife@telecom.tuc.gr, apotam@gmail.com

Abstract

We describe the grammar induction system for Spoken Dialogue Systems (SDS) submitted to SemEval'14: Task 2. A statistical model is trained with a rich feature set and used for the selection of candidate rule fragments. Posterior probabilities produced by the fragment selection model are fused with estimates of phrase-level similarity based on lexical and contextual information. Domain and language portability are among the advantages of the proposed system that was experimentally validated for three thematically different domains in two languages.

1 Introduction

A critical task for Spoken Dialogue Systems (SDS) is the understanding of the transcribed user input, that utilizes an underlying domain grammar. An obstacle to the rapid deployment of SDS to new domains and languages is the time-consuming development of grammars that require human expertise. Machine-assisted grammar induction has been an open research area for decades (K. Lari and S. Young, 1990; S. F. Chen, 1995) aiming to lower this barrier. Induction algorithms can be broadly distinguished into resource-based, e.g., (A. Ranta, 2004), and data-driven, e.g., (H. Meng and K.-C. Siu, 2002). The main drawback of the resource-based paradigm is the requirement of pre-existing knowledge bases. This is addressed by the data-driven paradigm that relies (mostly) on plain corpora. SDS grammars are built by utilizing low- and high-level rules. Low-level rules are similar to gazetteers consisting of terminal entries, e.g., list of city names. High-level rules can be lexicalized as textual fragments (or chunks), which are semantically defined on top of low-level rules, e.g., `'depart from <City>'`. The data-driven induction of low-level rules is a well-researched area enabled by various technologies including web harvesting for corpora creation (Klasinas et al., 2013), term extraction (K. Frantzi and S. Ananiadou, 1997), word-level similarity computation (Pargellis et al., 2004) and clustering (E. Iosif and A. Potamianos, 2007). High-level rule induction is a less researched area that poses two main challenges: 1) the extraction and selection of salient candidate fragments from a corpus that convey semantics relevant to the domain of interests and 2) the organization of such fragments (e.g., via clustering) according to their semantic similarity. Despite the recent interest on phrase (J. Mitchell and M. Lapata, 2010) and sentence similarity, each respective problem remains open.

Next, our submission[1] for the SemEval'14: Task2 is briefly described, which constitutes a data-driven approach for inducing high-level SDS grammar rules. At the system's core lies a statistical model for the selection of textual fragments based on a rich set of features. This set includes various lexical features, augmented with statistics from n-gram language models, as well as with heuristic features. The candidate selection model posterior is fused with a phrase-level semantic similarity metric. Two different approaches are used for similarity computation relying on the overlap of character bigrams or context-based similarity according to the distributional hypothesis of meaning. The domain and language portability of the proposed system is demonstrated by its successful application across three different domains and

[1]Please note that the last three authors of this submission are among the organizers of this task.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 668–672, Dublin, Ireland, August 23-24, 2014.

two languages. All the four subtasks defined by the organizers were completed with very good performance that exceeds the baseline.

2 System Description

The basic functionality of the proposed system is the mapping (assignment) of unknown textual fragments into known high-level grammar rules. Let E be the set of unknown fragments, while the set of known rules is denoted by R. Each unknown fragment $f \in E$ is allowed to be mapped to a single high-level rule $r_s \in R$, where $1 \leq s \leq m$ and m is the total number of rules in the grammar.

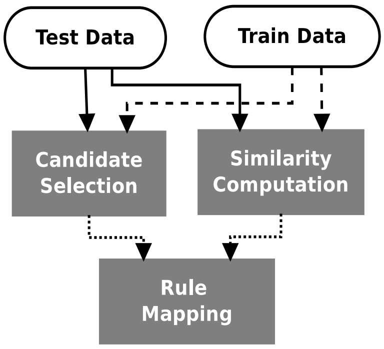

Figure 1: Overview of system architecture.

The system consists of three major components as shown at the system architecture diagram in Fig. 1, specifically: 1) candidate selection: a set of classifiers is built, one for each r_s to select whether $f \in E$ is a candidate member of the specific rule[2], 2) similarity computation between f and r_s, and 3) mapping f to a high-level rule r_s (denoted as $f \mapsto r_s$) according to the following model:

$$\underset{s}{\operatorname{argmax}}\{p(r_s|f)^w S(f, r_s)\} : f \mapsto r_s \qquad (1)$$

where $p(r_s|f)$ stands for the probability of f belonging to rule r_s and it is estimated via the respective classifier. The similarity between f and r_s is denoted by $S(f|r_s)$, while w is a fixed weight taking values in the interval $[0\ \infty)$. The fusion weight w controls the relative importance of the candidate selection and semantic similarity modules, e.g., for $w = 0$ only the similarity metric $S(f, r_s)$ is used in the decision. For example, consider the fragment f 'leaving <City>'. Also, assume two high-level rules, namely, <ArrCity>={'arrive at <City>',...} and <DepCity>= {'depart <City>',...}. According to (1) f is mapped to the <DepCity> rule.

[2]The requirement for building a classifier for each grammar rule is realistic for the case of SDS, especially for the typical iterative human-in-the-loop grammar development scenario.

2.1 Candidate Selection

In this section, the features used for building the candidate selection module for each $r_s \in R$ are briefly described. Given a pair (f,r_s) a two-class statistical classification model that corresponds to r_s is used for estimating $p(r_s|f)$ in (1).

Definitions. A high-level rule r_s can be considered as a set of fragments, e.g.,'depart <City>', 'leaving <City>'. For each fragment there are two types of constituents, namely, lexical (e.g., 'depart','leaving') and low-level rules (e.g., '<City>'). The following features are extracted for r_s considering its respective fragments, as well as for f.

Shallow features. 1) the number of constituents (i.e., tokens), 2) the count of lexical constituents to the number of tokens, 3) the count of low-level rules to the number of tokens, 4) the count of lexical constituents that follow the right-most low-level rule of the fragment, and 5) the count of low-level rules that appear twice in a fragment.

Perplexity-based features. A fragment $\tilde{f}$ can be represented as a sequence of tokens as $w_1\ w_2\ ...\ w_z$. The perplexity of $\tilde{f}$ is defined as $PP(\tilde{f}) = 2^{H(\tilde{f})}$, where $H(\tilde{f}) = \frac{1}{z}\log(p(\tilde{f}))$. $p(\tilde{f})$ stands for the probability of $\tilde{f}$ estimated using an n-gram language model. Two PP values were used as features computed for $n = 2, 3$.

Features of lexical similarity. Four scores of lexical similarity computed between f and r_s were used as features. Let N_s denote the set of fragments that are included in the training set of each rule r_s. The following metrics were employed for computing the similarity between the unknown fragment f and a fragment $f_s \in N_s$: 1) the normalized longest common subsequence (Stoilos et al., 2005) denoted as S_C, 2) the normalized overlap in character bigrams that is denoted as S_B and it is defined in (2), 3) a proposed variation of the Levenshtein distance, S_L, defined as $S_L(f, f_s) = \frac{l_1 - L(f,f_s)}{l_1 + d}$, where l_1 and l_2 are the lengths (in characters) of the lengthiest and the shortest fragment between f and f_s, respectively, while $d = l_1 - l_2$. $L(.)$ stands for the Levenshtein distance (V. I. Levenshtein, 1966; R. A. Wagner and M. J. Fischer, 1974). 4) if f and f_s differ by one token exactly S_L is applied, otherwise their similarity is set to 0. Regarding S_C and S_B, the similarity between

f and r_s was estimated as the maximum similarity yielded when computing the similarities between f and each $f_s \in N_s$. For the rest metrics, the similarity between f and r_s was estimated by averaging the $|N_s|$ similarities computed between f and each $f_s \in N_s$.

Heuristic features. Considering an unknown fragment f and the set of training fragments N_s corresponding to rule r_s, in total nine features were used: 1) the difference between the average length (in tokens) of fragments in N_s and the length of f, 2) the difference between the average number of low-level rules in N_s and the number of low-level rules in f, 3) as 2) but considering the lexical constituents instead of low-level rules, 4) the number of low-level rules shared between N_s and f, 5) as 4) but considering the lexical constituents instead of low-level rules, 6) a boolean function that equals 1 if f is a substring of at least one $f_s \in N_s$, 7) a boolean function that equals 1 if f shares the same lexical constituents at least one $f_s \in N_s$, 8) a boolean function that equals 1 if f is shorter by one token compared to any $f_s \in N_s$, 9) a boolean function that equals 1 if f is lengthier by one token compared to any $f_s \in N_s$.

Selection. The aforementioned features are used for building a binary classifier for each $r_s \in R$, where $1 \leq s \leq m$, for deciding whether f can be regarded as a candidate member of r_s or not. Given an unknown fragment f these classifiers are employed for estimating in total m probabilities $p(r_s|f)$.

2.2 Similarity Metrics

Here, two types of similarity metrics are defined, which are used for estimating $S(f, r_s)$ in (1).

String-based similarity. Consider two fragments f_i and f_j whose sets of character bigrams are denoted as M_i and M_j, respectively. Also, $M_{min} = \min(|M_i|, |M_j|)$ and $M_{max} = \max(|M_i|, |M_j|)$. The similarity between f_i and f_j is based on the overlap of their respective character bigrams defined as (Jimenez et al., 2012):

$$S_B(f_i, f_j) = \frac{|M_i \cap M_j|}{\alpha M_{max} + (1-\alpha) M_{min}}, \quad (2)$$

where $0 \leq \alpha \leq 1$, while, here we use $\alpha = 0.5$. The similarity between a fragment f and a rule r_s is computed by averaging the similarities computed between f and each $f_s \in N_s$.

Context-based similarity. This is a corpus-based metric relying on the distributional hypothesis of meaning suggesting that *similarity of context implies similarity of meaning* (Z. Harris, 1954). A contextual window of size $2K+1$ words is centered on the fragment of interest f_i and lexical features are extracted. For every instance of f_i in the corpus the K words left and right of f_i formulate a feature vector v_i. For a given value of K the context-based semantic similarity between two fragments, f_i and f_j, is computed as the cosine of their feature vectors: $S^K(f_i, f_j) = \frac{v_i \cdot v_j}{||v_i||\,||v_j||}$. The elements of feature vectors can be weighted according various schemes (E. Iosif and A. Potamianos, 2010), while, here we use a binary scheme. The similarity between a fragment f and a rule r_s is computed by averaging the similarities computed between f and each $f_s \in N_s$.

2.3 Mapping of Unknown Fragments

The output of the described system is the mapping of a fragment f to a single (i.e., one-to-one assignment) high-level rule $r_s \in R$, where $1 \leq s \leq m$. This is achieved by applying (1). The $p(r_s|f)$ probabilities were estimated as described in Section 2.1. The $S(f, r_s)$ similarities were estimated using either S^K or S_B defined in Section 2.2.

3 Datasets and Experiments

Datasets. The data was organized with respect to three different domains: 1) air travel (flight booking, car rental etc.), 2) tourism (information for city guide), and 3) finance (currency exchange). In total, there are four separate datasets: two datasets for the air travel domain in English (EN) and Greek (GR), one dataset for the tourism domain in English, and one dataset for the finance domain in English.

The number of high-level rules for each dataset are shown in Table 1, along with the number of fragments included in training and test data.

Domain	#rules	#train frag.	#test frag.
Travel:EN	32	982	284
Travel:GR	35	956	324
Tourism:EN	24	1004	285
Finance:EN	9	136	37

Table 1: Number of rules and train/test fragments.

Experiments. Regarding the computation of perplexity-based features (defined in Section 2.1) the SRILM toolkit (A. Stolcke, 2002) was used. The n-gram probabilities were estimated over a corpus that was created by aggregating all the valid fragments included in the training data. For the computation of the context-based similarity metric S^K (defined in Section 2.2) a corpus of web-harvested data was created for each domain/language. The context window size K was set to 1. The size of the used corpora are presented Table 2, while the process of corpus creation is detailed in (Klasinas et al., 2013). The classifiers used for the candidate selection module, described in Section 2.1 were random forests with 50 trees (L. Breiman, 2001).

Domain	# sentences
Travel:EN	5721
Travel:GR	6359
Tourism:EN	829516
Finance:EN	168380

Table 2: Size of corpora used in S^K metric.

4 Evaluation Metrics and Results

The proposed model defined by (1) was evaluated in terms of weighted F-measure, (FM). Initially, we run our system using the training and development set provided by the task organizers, in order to tune the w and K parameters. The tuning was conducted on the Travel English domain, while the respective evaluation results are shown in Table 3 in terms of FM. We observe that the best results are achieved for $w = 1$ and $w = 500$. In the case where $w = 0$ the rule mapping relies only on the similarity metric. In addition, we experimented with various values the context window size K of the context-based similarity metric S^K: $K = 1, 3, 7$. For all values of K similar performance was obtained (0.70). Given the aforementioned tuning the following values were selected for the official runs: $w = 1$, $w = 500$ and $K = 1$. In total, three system runs were submitted:

Weight w	0	1	50	500
FM	0.68	0.72	0.70	0.72

Table 3: Results for the tuning of w.

Domains	Baseline	Run 1	Run 2	Run 3
Travel:EN	0.51	0.66	0.65	**0.68**
Travel:GR	0.26	**0.52**	0.49	0.49
Tourism:EN	**0.87**	0.86	0.85	0.86
Finance:EN	0.60	**0.70**	0.63	0.58
UA	0.56	**0.69**	0.66	0.65
WA	0.52	**0.66**	0.64	0.65

Table 4: Official results.

Run 1. The character bigram similarity metric was used, while w was set to 1.
Run 2. The context-based similarity metrics was used with $K = 1$, while w was set to 1.
Run 3. The character bigram similarity metric was used, while w was set to 500.
The results for the aforementioned runs, along with the baseline performance are shown in Table 4. An overview of the participating systems suggests that our submission achieved the highest performance for almost all domains and languages. The weighted (WA) and unweighted (UA) average across the 4 datasets are also presented, where the weight depends on the number of rules in the dataset. Using these measures, our main run (Run 1) obtained the best results. We observe that the performance is consistently worse for Runs 2 and 3, with the exception of the Travel English dataset. Comparing the performance of Runs 1 and 2, we observe that the character bigram metric consistently outperforms the context-based one. For individual datasets, our system underperforms for the Finance (in Run 3) and the Tourism domain (in all Runs). For the case of the Finance domain this may be attributed to the relatively limited training data.

5 Conclusions

We proposed a supervised grammar induction system using the fusion of a grammar fragment selection and similarity estimation modules. The best configuration of our system was Run 1 which achieved the highest performance compared to other submissions, in almost all domains. To summarize, 1) the selection module boost the system's performance significanlty, 2) the high performance in different domains is a promising indicator for domain and language portability. Future work should involve the implementation of more complex features for the candidate selection algorithm and further investigation of phrase level similarity metrics.

Acknowledgements

This work has been partially funded by the projects: 1) SpeDial, and 2) PortDial, supported by the EU Seventh Framework Programme (FP7), with grant number 611396 and 296170, respectively.

References

Elias Iosif and Alexandros Potamianos. 2010. *Unsupervised semantic similarity computation between terms using web documents.* IEEE Transactions on Knowledge and Data Engineering, 22(11), pp. 1637-1647.

Sergio Jimenez, Claudia Becerra and Alexander Gelbukh. 2012. *Soft Cardinality: A parameterized similarity function for text comparison.* In Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM), pp. 449-453

Ioannis Klasinas, Alexandros Potamianos, Elias Iosif, Spyros Georgiladakis and Gianluka Mameli. 2013. *Web data harvesting for speech understanding grammar induction.* in Proceedings of the Interspeech.

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), pp. 172-181.

PortDial Project free data deliverable D3.1. https://sites.google.com/site/portdial2/deliverables-publication

Andreas Stolcke 2002 *Srilm-an extensible language modeling toolkit* in Proceedings of the Interspeech 2002

Karim Lari and Steve J. Young 2002. *The estimation of stochastic context-free grammars using the inside-outside algorithm.* Computer Speech and Language, 4(1), pp. 35-56.

Stanley F. Chen 1995. *Bayesian grammar induction for language modeling.* in Proceedings of the 33rd annual meeting of ACL

Zellig Harris 1954. *Distributional structure.* Word, 10(23), pp. 146-162.

Rebecca Hwa 1999. *Supervised grammar induction using training data with limited constituent information.* in Proceedings of the 37th annual meeting of ACL

Matthew Lease, Eugene Charniak, and Mark Johnson 2005. *Parsing and its applications for conversational speech.* in Proceedings of Acoustics, Speech, and Signal Processing (ICASSP)

Vladimir I. Levenshtein 1966. *Binary codes capable of correcting deletions, insertions and reversals.* in Soviet physics doklady, 10(8), pp. 707-710.

Leo Breiman 2001. *Random forests.* in Machine Learning, 45(1), pp. 5-32.

Dan Jurafsky and James H. Martin 2009. *Speech and language processing an introduction to natural language processing, computational linguistics, and speech.* Pearson Education Inc

Giorgos Stoilos, Giorgos Stamou, and Stefanos Kollias 2005. *A string metric for ontology alignment.* in The Semantic WebISWC, pp. 624637

Robert A. Wagner and Michael J. Fisher 1974. *The string-to-string correction problem.* Journal of the ACM (JACM), 21(1), pp. 168-173

Katerina Frantzi and Sophia Ananiadou 1997. *Automatic term recognition using contextual cues.* in Proceedings of International Joint Conferences on Artificial Intelligence

Elias Iosif and Alexandros Potamianos 2007. *A soft-clustering algorithm for automatic induction of semantic classes.* in Proceedings of Interspeech

Jeffrey Mitchell and Mirela Lapata 2010. *Composition in distributional models of semantics.* Cognitive Science, 34(8):1388-1429.

Ye-Yi Wang and Alex Acero 2006. *Rapid development of spoken language understanding grammars.* Speech Communication, 48(3), pp. 360-416.

Eric Brill 1992. *A simple rule-based part of speech tagger.* in Proceedings of the workshop on Speech and Natural Language

Alexander Clark 2001. *Unsupervised induction of stochastic context-free grammars using distributional clustering.* in Proceedings of the 2001 workshop on Computational Natural Language Learning

Benjamin Snyder, Tahira Naseem, and Regina Barzilay 2009. *Unsupervised multilingual grammar induction.* in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL

Aarne Ranta 2009. *Grammatical framework: A type-theoretical grammar formalism.* Journal of Functional Programming: 14(2), pp. 145-189

Andrew Pargellis, Eric Fosler-Lussier, Chin Hui Lee, Alexandros Potamianos and Augustine Tsai 2009. *Auto-induced Semantic Classes.* Speech Communication: 43(3), pp. 183-203

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre 2012. *SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity.* in Proceedings of the First Joint Conference on Lexical and Computational Semantics (*Sem), pp. 385-393