Adobe-MIT submission to the DSTC 4 Spoken Language Understanding pilot task

Franck Dernoncourt, Ji Young Lee, Trung H. Bui, and Hung H. Bui

Abstract
The Dialog State Tracking Challenge 4 (DSTC 4) proposes several pilot tasks. In this paper, we focus on the spoken language understanding pilot task, which consists of tagging a given utterance with speech acts and semantic slots. We compare different classifiers: the best system obtains 0.52 and 0.67 F1-scores on the test set for speech act recognition for the tourist and the guide respectively, and 0.52 F1-score for semantic tagging for both the guide and the tourist.

1 Speech act recognition

Recognizing the speech acts of the current utterance is one of the two goals of the spoken language understanding pilot task. In the training and development sets, each utterance is annotated with one speech act. One speech act is composed of zero, one or two speech act categories. Each speech act category has in turn zero, one or two speech act attributes. There are 4 speech act categories, and 22 speech act attributes. [6] and [7] give further details on the task. The main approaches for this task are presented in [15, 1, 17, 5, 16, 19, 10, 3].

We submitted 5 systems. Systems 3 and 5 were the best performing ones. System 3 is based on a support vector machine (SVM) classifier to recognize the speech acts: the features are the 5000 most common unigrams, bigrams, trigrams, as well as a binary feature indicating whether the current speaker is different from the speaker in the last utterance. To account for the history, each feature is computed for both the current and the previous utterance. Two SVM classifiers were trained: one for each speaker. The kernel function as well as the penalty parameter of the error term were both optimized with 5-fold cross-validation. System 5 is similar, but with logistic regression as the classifier; moreover, it uses one single speaker-independent model instead of one model per speaker, as it slightly improves the results on the development set. Systems 3 and 5 assume that each utterance contains exactly one speech act

Franck Dernoncourt
Adobe Research, San Jose, CA, USA and MIT, Cambridge, MA, USA e-mail: francky@mit.edu

Ji Young Lee
Massachusetts Institute of Technology, Cambridge, MA, USA e-mail: jjylee@mit.edu

Trung H. Bui
Adobe Research, San Jose, CA, USA e-mail: bui@adobe.com

Hung H. Bui
Adobe Research, San Jose, CA, USA e-mail: hubui@adobe.com

category and one speech act attribute: they are therefore multiclass, monolabel classifiers, with 88 possible classes (4 speech act categories $\times$ 22 speech act attributes).

System 4 is based on a random forest classifier and has only 4 features: the number of question marks (discrete value), whether the current speaker is different from the speaker in the last utterance, whether the current speaker is different from the speaker in the second to previous utterance, and whether the current speaker is the guide or the tourist. System 4 was designed to predict the speech act categories, but not the speech act attributes. System 2 is the same as System 4, except that System 4's features are computed on the current and previous utterances, while System 2's features are computed on the current, previous and second-to-previous utterances.

System 1 is a rule-based classifier consisting of set of around 10 simple rules (e.g. if the preceding utterance is predicted as a question, then the current utterance is a response): it was designed to be used as a baseline. Table 1 presents the results.

2 Semantic tagging

Semantic tagging is the second goal of the SLU pilot task. A tagged entity comprises one or several words. A tag includes one of 8 main categories, and may contain a subcategory, a relative modifier, and a from-to modifier. The ontology contains the list of subcategories, relative modifiers, and from-to modifiers that are present in each main category. [6] and [7] give further details on the task. The main approaches for this task are presented in [8, 9, 14, 12, 18, 4, 11, 2].

Our semantic tagging system is based on conditional random fields (CRFs) implemented by the CRFsuite library [13] and uses the following features computed on 7 consecutive words (the current word, the 3 previous words, and the 3 following words): case-insensitive unigrams, the last 3 characters of the word, whether the first letter of the word is an uppercase, whether all the letters of the word are uppercases, whether the word contains a digit, the coarse-grained part-of-speech of the word, and the fine-grained part-of-speech of the word. Four CRFs are trained independently, one for each of the 4 types of attributes: main category, subcategory, relation, and from-to. To combine the output of each CRF, a semantic tag is first generated for each sequence of words tagged by the main category CRF. The other thee attributes are included in the semantic tag if these words are tagged by the corresponding CRFs with a value that is present in the main category according to the ontology. Table 1 presents the results.

Table 1 Results of different systems on the test set, evaluated by DSTC 4's organizers.

Tracker	Guide			Tourist		
	Precision	Recall	F1-score	Precision	Recall	F1-score
System 1	0.6287	0.5191	0.5687	0.3583	0.2977	0.3252
System 2	0.6330	0.5227	0.5726	0.2931	0.2435	0.2660
System 3	**0.7451**	**0.6153**	**0.6740**	0.5627	0.4675	0.5107
System 4	0.6314	0.5214	0.5712	0.2939	0.2442	0.2668
System 5	0.6762	0.5584	0.6117	**0.5736**	**0.4766**	**0.5206**
Semantic	0.5646	0.4886	0.5239	0.5741	0.4764	0.5207

Acknowledgements The authors would like to warmly thank the DSTC 4 team for organizing the challenge and being so prompt to respond to emails. The authors are also grateful to the anonymous reviewers as well as to Walter Chang for their valuable feedback.

References

1. J. Ang, Y. Liu, and E. Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. In *ICASSP (1)*, pages 1061–1064, 2005.
2. L. Deng, G. Tur, X. He, and D. Hakkani-Tur. Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 210–215. IEEE, 2012.
3. M. Fišel. Machine learning techniques in dialogue act recognition. *Eesti Rakenduslingvistika Ühingu aastaraamat*, (3):117–134, 2007.
4. D. Guo, G. Tur, W.-t. Yih, and G. Zweig. Joint semantic utterance classification and slot filling with recursive neural networks. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 554–559. IEEE, 2014.
5. G. Ji and J. Bilmes. Dialog act tagging using graphical models. In *ICASSP (1)*, pages 33–36, 2005.
6. S. Kim, L. F. D'Haro, R. E. Banchs, J. Williams, and M. Henderson. Dialog State Tracking Challenge 4: Handbook, 2015.
7. S. Kim, L. F. D'Haro, R. E. Banchs, J. Williams, and M. Henderson. The Fourth Dialog State Tracking Challenge. In *Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS)*, 2016.
8. T. Kudo and Y. Matsumoto. Chunking with support vector machines. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.
9. J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
10. L. Levin, K. Ries, A. Thyme-Gobbel, and A. Lavie. Tagging of speech acts and dialogue games in spanish call home. In *Workshop: Towards Standards and Tools for Discourse Tagging*, pages 42–47, 1999.
11. G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, et al. Using recurrent neural networks for slot filling in spoken language understanding. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23(3):530–539, 2015.
12. G. Mesnil, X. He, L. Deng, and Y. Bengio. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, pages 3771–3775, 2013.
13. N. Okazaki. CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007.
14. C. Raymond and G. Riccardi. Generative and discriminative algorithms for spoken language understanding. In *INTERSPEECH*, pages 1605–1608, 2007.
15. K. Ries. Hmm and neural network based speech act detection. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 1, pages 497–500. IEEE, 1999.
16. R. Serafin and B. Di Eugenio. Flsa: Extending latent semantic analysis with features for dialogue act classification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 692. Association for Computational Linguistics, 2004.
17. A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.
18. K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi. Spoken language understanding using long short-term memory neural networks. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 189–194. IEEE, 2014.
19. M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke. *Toward joint segmentation and classification of dialog acts in multiparty meetings*. Springer, 2006.