Using Semantic Features to Improve Task Identification in Email Messages

Shinjae Yoo[1], Donna Gates[1], Lori Levin[1], Simon Fung[1],
Sachin Agarwal[1], and Michael Freed[2]

[1] Language Technology Institute
5000 Forbes Ave, Pittsburgh, PA 15213
{sjyoo,dmg,lsl,sfung,sachina}@cs.cmu.edu
[2] SRI International
333 Ravenswood Ave., Menlo Park, CA 94025
freed@ai.sri.com

Abstract. Automated identification of tasks in email messages can be very useful to busy email users. What constitutes a task varies across individuals and must be learned for each user. However, training data for this purpose tends to be scarce. This paper addresses the lack of training data using domain-specific semantic features in document representation for reducing vocabulary mismatches and enhancing the discriminative power of trained classifiers when the number of training examples is relatively small.

Keywords: classification, semantic features, construction grammar, ontology.

1 Introduction

Lack of training data and idiosyncratic, highly ambiguous user definitions of tasks make email task classification very challenging. A simple bag-of-words approach is problematic mainly due to vocabulary mismatches. Some such mismatches can be resolved by proper text preprocessing such as stemming or tokenization. However, the differentiation of word senses and the generalization of word meanings for synonyms are not easily solved. Bloehdorn and Hortho [2] explored the use of ontologies (semantic hierarchies) such as WordNet and claimed that classification performance improved on a standard benchmark dataset such as Reuters 21578. However, their baseline performance was lower than the baseline reported by Li and Yang [2] where no such additional Natural Language Processing features were used.

Our hypothesis is that classification performance can be improved by using domain specific semantic features. In this paper, we investigate the use of semantic features to overcome the vocabulary mismatch problem using a general ontology and domain specific knowledge engineering. When we extracted the ontology information using domain specific grammars indicating the appropriate semantic ancestor (e.g., *conference-session* versus the too general *thing* or too specific *tutorial* concepts), we observed improvement over naively extracting ontology information.

E. Kapetanios, V. Sugumaran, M. Spiliopoulou (Eds.): NLDB 2008, LNCS 5039, pp. 355–357, 2008.

2 Natural Language Feature Extraction

Our approach uses rules to extract features from natural language text. We use syntactic information from dependency parses produced by Connexor Machinese Syntax[4] and semantic concepts from the RADAR's Scone[6] ontology system. Sets of construction rules (inspired by Construction Grammar [3]) map Connexor's syntactic structures onto semantic frames. The construction rules encode patterns that are matched against the syntactic parse to add general pragmatic information (e,g., speech-acts: *request-information, request-action, reject-proposal*), semantic information (e.g., semantic roles: *agent, patient, location*) and domain specific features (e.g., *conference-session, briefing, meeting*) to the final semantic frame. The inventory of domain specific features was determined manually. At run time, features are extracted using ConAn rules that combine Scone concepts with speech-act information and/or semantic roles. These domain features are tied to domain specific classes or sub-tasks (e.g., *information, web, meetings, food, conference sessions,* etc…) identified by the classifier.

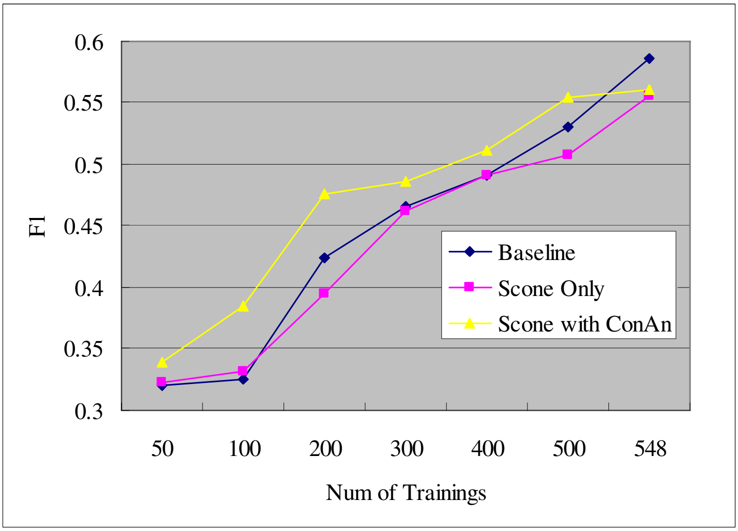

Fig. 1. Macro F1 of SVM

3 Evaluation and Results

We tested our semantic features using SVM [1] on 548 training emails and evaluated on 107 unseen emails. The training emails were hand-labeled with eight previously defined categories (task types). We tested two conditions: first, naively extracting Scone concepts for each word and its semantic parent(s) and grandparent(s) and then, second, extracting semantic features to find the most appropriate domain specific ancestor(s) using construction rules. Naively extracted Scone concepts hurt system performance. However, the construction grammar driven semantic features did well on training data of up to 500 emails. The ontology, construction analyzer and classifier modules described in this paper are being developed as a part of RADAR [5], a system that will act as a personal assistant in such tasks as classifying e-mail, updating web information, scheduling rooms and scheduling meetings.

Acknowledgement

The material described in this paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

1. Li, F., Yang, Y.: A loss function analysis for classification methods in text categorization. In: ICML 2003 (2003)
2. Bloehdorn, S., Hortho, A.: Boosting for Text Classification with Semantic Features. In: Mobasher, B., Nasraoui, O., Liu, B., Masand, B. (eds.) WebKDD 2004. LNCS (LNAI), vol. 3932, pp. 149–166. Springer, Heidelberg (2006)
3. Goldberg, A.E.: Constructions: a new theoretical approach to language. Trends in Cognitive Sciences 7(5), 219–224 (2003)
4. Tapanianen, P., Järvinen, T.: A non projective dependency parser. In: Fifth Conference on Applied Natural Language Processing, pp. 64–71 (1997)
5. Freed, M., Carbonell, J., Gordon, G., Hayes, J., Myers, B., Siewiorek, D., Smith, S., Steinfeld, A., Tomasic, A.: RADAR: A Personal Assistant That Learns to Reduce Email Overload. In: AAAI 2008 (2008)
6. Fahlmann, S.: Scone User's Manual (2006), `http://www.cs.cmu.edu/~sef/scone/`