Domain Adaptation with Artificial Data for Semantic Parsing of Speech

Lonneke van der Plas
Department of Linguistics
University of Geneva
Geneva, Switzerland

James Henderson
Department of Computer Science
University of Geneva
Geneva, Switzerland

Paola Merlo
Department of Linguistics
University of Geneva
Geneva, Switzerland

{Lonneke.vanderPlas,James.Henderson,Paola.Merlo}@unige.ch

Abstract

We adapt a semantic role parser to the domain of goal-directed speech by creating an artificial treebank from an existing text treebank. We use a three-component model that includes distributional models from both target and source domains. We show that we improve the parser's performance on utterances collected from human-machine dialogues by training on the artificially created data without loss of performance on the text treebank.

1 Introduction

As the quality of natural language parsing improves and the sophistication of natural language understanding applications increases, there are several domains where parsing, and especially semantic parsing, could be useful. This is particularly true in adaptive systems for spoken language understanding, where complex utterances need to be translated into shallow semantic representation, such as dialogue acts.

The domain on which we are working is goal-directed system-driven dialogues, where a system helps the user to fulfil a certain goal, e.g. booking a hotel room. Typically, users respond with short answers to questions posed by the system. For example *In the South* is an answer to the question *Where would you like the hotel to be?* Parsing helps identifying the components (*In the South* is a PP) and semantic roles identify the PP as a locative, yielding the following slot-value pair for the dialogue act: *area=South*. A PP such as *in time* is not identified as a locative, whereas keyword-spotting techniques as those currently used in dialogue systems may produce *area=South* and *area=time* indifferently.

Statistical syntactic and semantic parsers need treebanks. Current available data is lacking in one or more respects: Syntactic/semantic treebanks are developed on text, while treebanks of speech corpora are not semantically annotated (e.g. Switchboard). Moreover, the available human-human speech treebanks do not exhibit the same properties as the system-driven speech on which we are focusing, in particular in their proportion of non-sentential utterances (NSUs), utterances that are not full sentences. In a corpus study of a subset of the human-human dialogues in the BNC, Fernández (2006) found that only 9% of the total utterances are NSUs, whereas we find 44% in our system-driven data.

We illustrate a technique to adapt an existing semantic parser trained on merged Penn Treebank/PropBank data to goal-directed system-driven dialogue by artificial data generation. Our main contribution lies in the framework used to generate artificial data for domain adaptation. We mimic the distributions over parse structures in the target domain by combining the text treebank data and the artificially created NSUs, using a three-component model. The first component is a hand-crafted model of NSUs. The second component describes the distribution over full sentences and types of NSUs as found in a minimally annotated subset of the target domain. The third component describes the distribution over the internal parse structure of the generated data and is taken from the source domain.

Our approach differs from most approaches to domain adaptation, which require some training on fully annotated target data (Nivre et al., 2007), whereas we use minimally annotated target data only to help determine the distributions in the artificially created data. It also differs from previ-

Proceedings of NAACL HLT 2009: Short Papers, pages 125–128,
Boulder, Colorado, June 2009.

ous work in domain adaptation by Foster (2007), where similar proportions of ungrammatical and grammatical data are combined to train a parser on ungrammatical written text, and by Weilhammer et al. (2006), who use interpolation between two separately trained models, one on an artificial corpus of user utterances generated by a hand-coded domain-specific grammar and one on available corpora. Whereas much previous work on parsing speech has focused on speech repairs, e.g. Charniak and Johnson (2001), we focus on parsing NSUs.

2 The first component: a model of NSUs

To construct a model of NSUs we studied a subset of the data under consideration: TownInfo. This small corpus of transcribed spoken human-machine dialogues in the domain of hotel/restaurant/bar search is gathered using the TownInfo tourist information system (Lemon et al., 2006).

The NSUs we find in our data are mainly of the type answers, according to the classification given in Fernández (2006). More specifically, we find short answers, plain and repeated affirmative answers, plain and helpful rejections, but also greetings.

Current linguistic theory provides several approaches to dealing with NSUs (Merchant, 2004; Progovac et al., 2006; Fernández, 2006). Following the linguistic analysis of NSUs as non-sentential small clauses (Progovac et al., 2006) that do not have tense or agreement functional nodes, we make the assumption that they are phrasal projections. Therefore, we reason, we can create an artificial data set of NSUs by extracting phrasal projections from an annotated treebank.

In the example given in the introduction, we saw a PP fragment, but fragments can be NPs, APs, etc. We define different types of NSUs based on the root label of the phrasal projection and define rules that allow us to extract NSUs (partial parse trees) from the source corpus.[1] Because the target corpus also contains full sentences, we allow full sentences to be taken without modification from the source treebank.

[1]Not all of these rules are simple extractions of phrasal projections, as described in section 4.

3 The two distributional components

The distributional model consists of two components. By applying the extraction rules to the source corpus we build a large collection of both full sentences and NSUs. The distributions in this collection follow the distributions of trees in the source domain (first distributional component). We then sample from this collection to generate our artificial corpus following distributions from the target domain (second distributional component).

The probability of an artificial tree $P(f_i(c_j))$ generated with an extraction rule f_i applied to a constituent from the source corpus c_j is defined as

$$P(f_i(c_j)) = P(f_i)P(c_j|f_i) \approx P_t(f_i)P_s(c_j|f_i)$$

The first distributional component originates from the source domain. It is responsible for the internal structure of the NSUs and full sentences extracted. $P_s(c_j|f_i)$ is the probability of the constituent taken from the source treebank (c_j), given that the rule f_i is applicable to that constituent.

Sampling is done according to distributions of NSUs and full sentences found in the target corpus ($P_t(f_i)$). As explained in section 2, there are several types of NSUs found in the target domain. This second component describes the distributions of types of NSUs (or full sentences) found in the target domain. It determines, for example, the proportion of NP NSUs that will be added to the artificial corpus.

To determine the target distribution we classified 171 (approximately 5%) randomly selected utterances from the TownInfo data, that were used as a development set.[2] In Table 1 we can see that 15.2 % of the trees in the artificial corpus will be NP NSUs.[3]

4 Data generation

We constructed our artificial corpus from sections 2 to 21 of the Wall Street Journal (WSJ) section of the Penn Treebank corpus (Marcus et al., 1993)

[2]We discarded very short utterances (yes, no, and greetings) since they don't need parsing. We also do not consider incomplete NSUs resulting from interruptions or recording problems.

[3]Because NSUs can be interpreted only in context, the same NSU can correspond to several syntactic categories: *South* for example, can be an noun, an adverb, or an adjective. In case of ambiguity, we divided the score up for the several possible tags. This accounts for the fractional counts.

Category	# Occ.	Perc.	Category	# Occ.	Perc.
NP	19.0	15.2	RB	1.7	1.3
JJ	12.7	10.1	DT	1.0	0.8
PP	12.0	9.6	CD	1.0	0.8
NN	11.7	9.3	Total frag.	70.0	56.0
VP	11.0	8.8	Full sents	55.0	44.0

Table 1: Distribution of types of NSUs and full sentences in the TownInfo development set.

merged with PropBank labels (Palmer et al., 2005). We included all the sentences from this dataset in our artificial corpus, giving us 39,832 full sentences. In accordance with the target distribution we added 50,699 NSUs extracted from the same dataset. We sampled NSUs according to the distribution given in Table 1. After the extraction we added a root FRAG node to the extracted NSUs[4] and we capitalised the first letter of each NSU to form an utterance.

There are two additional pre-processing steps. First, for some types of NSUs maximal projections are added. For example, in the subset from the target source we saw many occurrences of nouns without determiners, such as *Hotel* or *Bar*. These types of NSUs would be missed if we just extracted NPs from the source data, since we assume that NSUs are maximal projections. Therefore, we extracted single nouns as well and we added the NP phrasal projections to these nouns in the constructed trees. Second, not all extracted NSUs can keep their semantic roles. Extracting part of the sentence often severs the semantic role from the predicate of which it was originally an argument. An exception to this are VP NSUs and prepositional phrases that are modifiers, such as locative PPs, which are not dependent on the verb. Hence, we removed the semantic roles from the generated NSUs except for VPs and modifiers.

5 Experiments

We trained three parsing models on both the original non-augmented merged Penn Treebank/Propbank corpus and the artificially generated augmented treebank including NSUs. We ran a contrastive experiment to examine the usefulness of the three-component model by training two versions of the augmented model: One with and one without the target component.[5]

These models were tested on two test sets: a small corpus of 150 transcribed utterances taken from the TownInfo corpus, annotated with gold syntactic and semantic annotation by two of the authors[6]: the TownInfo test set. The second test set is used to compare the performance of the parser on WSJ-style sentences and consists of section 23 of the merged Penn Treebank/Propbank corpus. We will refer to this test set as the non-augmented test set.

5.1 The statistical parser

The parsing model is the one proposed in Merlo and Musillo (2008), which extends the syntactic parser of Henderson (2003) and Titov and Henderson (2007) with annotations which identify semantic role labels, and has competitive performance. The parser uses a generative history-based probability model for a binarised left-corner derivation. The probabilities of derivation decisions are modelled using the neural network approximation (Henderson, 2003) to a type of dynamic Bayesian Network called an Incremental Sigmoid Belief Network (ISBN) (Titov and Henderson, 2007).

The ISBN models the derivation history with a vector of binary latent variables. These latent variables learn to represent features of the parse history which are useful for making the current and subsequent derivation decisions. Induction of these features is biased towards features which are local in the parse tree, but can find features which are passed arbitrarily far through the tree. This flexible mechanism for feature induction allows the model to adapt to the parsing of NSUs without requiring any design changes or feature engineering.

5.2 Results

In Table 2, we report labelled constituent recall, precision, and F-measure for the three trained parsers (rows) on the two test sets (columns).[7] These mea-

[4]The node FRAG exists in the Penn Treebank. Our annotation does not introduce new labels, but only changes their distribution.

[5]The model without the target distribution has a uniform distribution over full sentences and NSUs and within NSUs a uniform distribution over the 8 types.

[6]This test set was constructed separately and is completely different from the development set used to determine the distributions in the target data.

[7]Statistical significance is determined using a stratified shuffling method, using software available at `http://www.cis.`

Training	Testing					
	TownInfo			PTB nonaug		
	Rec	Prec	F	Rec	Prec	F
PTB nonaug	69.4	76.7	72.9	81.4	82.1	81.7
PTB aug(+t)	81.4	77.8	79.5	81.3	82.0	81.7
PTB aug(−t)	62.6	64.3	63.4	81.2	81.9	81.6

Table 2: Recall, precision, and F-measure for the two test sets, trained on non-augmented data and data augmented with and without the target distribution component.

sures include both syntactic labels and semantic role labels.

The results in the first two lines of the columns headed *TownInfo* indicate the performance on the real data to which we are trying to adapt our parser: spoken data from human-machine dialogues. The parser does much better when trained on the augmented data. The differences between training on newspaper text and newspaper texts augmented with artificially created data are statistically significant ($p < 0.001$) and particularly large for recall: almost 12%.

The columns headed *PTB nonaug* show that the performance on parsing WSJ texts is not hurt by training on data augmented with artificially created NSUs (first vs. second line). The difference in performance compared to training on the non-augmented data is not statistically significant.

The last two rows of the TownInfo data show the results of our contrastive experiment. It is clear that the three-component model and in particular our careful characterisation of the target distribution is indispensable. The F-measure drops from 79.5% to 63.4% when we disregard the target distribution.

6 Conclusions

We have shown how a three-component model that consists of a model of the phenomenon being studied and two distributional components, one from the source data and one from the target data, allows one to create data artificially for training a semantic parser. Specifically, analysis and minimal annotation of only a small subset of utterances from the target domain of spoken dialogue systems suffices to determine a model of NSUs as well as the necessary target distribution. Following this framework we were able to improve the performance of a statistical parser on goal-directed spoken data extracted from human-machine dialogues without degrading the performance on full sentences.

upenn.edu/~dbikel/software.html.

Acknowledgements

The research leading to these results has received funding from the EU FP7 programme (FP7/2007-2013) under grant agreement nr 216594 (CLASSIC project: www.classic-project.org).

References

E. Charniak and M. Johnson. 2001. Edit detection and parsing for transcribed speech. In *Procs. NAACL*.

R. Fernández. 2006. *Non-sentential utterances in dialogue: classification resolution and use.* Ph.D. thesis, University of London.

J. Foster. 2007. Treebanks gone bad: Parser evaluation and retraining using a treebank of ungrammatical sentences. *International Journal of Document Analysis and Recognition*, 10:1–16.

J. Henderson. 2003. Inducing history representations for broad-coverage statistical parsing. In *Procs. NAACL-HLT*.

O. Lemon, K. Georgila, J. Henderson, and M. Stuttle. 2006. An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the TALK in-car system. In *Procs. EACL*.

M. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Comp. Ling.*, 19:313–330.

J. Merchant. 2004. Fragments and ellipsis. *Linguistics and Philosophy*, 27:661–738.

P. Merlo and G. Musillo. 2008. Semantic parsing for high-precision semantic role labelling. In *Procs. CONLL*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Procs. EMNLP-CoNLL*.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Comp. Ling.*, 31:71–105.

L. Progovac, K. Paesani, E. Casielles, and E. Barton. 2006. *The Syntax of Nonsententials:Multidisciplinary Perspectives*. John Benjamins.

I Titov and J Henderson. 2007. Constituent parsing with Incremental Sigmoid Belief Networks. In *Procs. ACL*.

K. Weilhammer, M. Stuttle, and S. Young. 2006. Bootstrapping language models for dialogue systems. In *Procs. Conf. on Spoken Language Processing*.