Cross-Lingual Spoken Language Understanding from Unaligned Data using Discriminative Classification Models and Machine Translation

Fabrice Lefèvre[1]*, François Mairesse*[2] *and Steve Young*[2]

[1]Université d'Avignon et des Pays de Vaucluse
Laboratoire Informatique d'Avignon (EA 931), F-84911 Avignon, France
[2]Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK
fabrice.lefevre@univ-avignon.fr, f.mairesse@eng.cam.ac.uk, sjy@cam.ac.uk

Abstract

This paper investigates several approaches to bootstrapping a new spoken language understanding (SLU) component in a *target* language given a large dataset of semantically-annotated utterances in some other *source* language. The aim is to reduce the cost associated with porting a spoken dialogue system from one language to another by minimising the amount of data required in the target language. Since word-level semantic annotations are costly, Semantic Tuple Classifiers (STCs) are used in conjunction with statistical machine translation models both of which are trained from *unaligned* data to further reduce development time. The paper presents experiments in which a French SLU component in the tourist information domain is bootstrapped from English data. Results show that training STCs on automatically translated data produced the best performance for predicting the utterance's dialogue act type, however individual slot/value pairs are best predicted by training STCs on the source language and using them to decode translated utterances.

Index Terms: spoken dialogue system, spoken language understanding, portability, bootstrapping

1. Introduction

Recent work has shown that statistical approaches to spoken language understanding (SLU) can be used to train models which can predict the meaning of unseen user utterances from a set of semantically-annotated training utterances [1, 2, 3, 5, 6]. The reduction of development time and cost is often cited as one of the main advantages of data-driven methods. However, developing a statistical SLU module for a new language remains costly, as it typically requires a large data collection phase. In this work we propose to circumvent this by porting an existing module from a *source* language to a *target* language, assuming that a large set of semantically-annotated utterances are available in the source language. As a consequence of recent breakthroughs in the field of statistical machine translation (SMT), translation systems can now be built at a reasonable development cost. We therefore propose to combine a state-of-the art SMT system with a data-driven SLU method to bootstrap an SLU component in the target language, while maintaining a high semantic accuracy.

Although keyword spotting techniques for SLU are robust to noise, they are generally too simple to model long-range dependencies within an utterance or to handle complex semantic representations (e.g., semantic trees). Research on SLU attempts to alleviate these issues by learning to derive a semantic representation from data. So far most work has focused on generative dynamic Bayesian networks that model the semantics of the utterance as a hidden structure on which observed words are conditioned [1, 2, 3, 4]. Such models can be trained on unaligned data, using expectation-maximisation techniques. But the Markovian assumption prevents such techniques from explicitly modelling long-range time dependencies. On the other hand, discriminative models do not make independence assumptions over the feature set and this can lead to improved performance. For instance linear-chain conditional random fields (CRF) have been shown to produce the best results when converting the SLU problem into a flat sequential labelling task [5]. However, a disadvantage of most discriminative methods is that they require the training utterances to be semantically annotated at the word-level. Aligning the semantic representation with individual words is a time-consuming task, which results in additional development cost when porting a dialogue system to a new language. In this context, the Semantic Tuple Classifiers (STC) approach has been introduced as an efficient yet simple technique that learns discriminative semantic concept classifiers *without requiring any alignment information* [6].

This paper proposes and evaluates several ways to efficiently use SMT systems to port an SLU module to a new language while minimising data annotation cost. The next section presents the SLU and SMT modules used in our experiments, as well as the different ways in which they can be combined for cross-lingual SLU. Section 4 details our experimental results in the tourist information domain, with English as the source language and French as the target language. Finally, Section 5 concludes with a discussion on future work.

2. Cross-lingual SLU using SMT

Cross-lingual SLU can be achieved in a variety of ways. Firstly, SMT can be used to translate the training data from the source to the target language. The translated data can then be used to train an SLU component in the target language. This method is referred to as *TrainOnTarget* in this paper. Although the training data is likely to include errors due to translation inaccuracy, this method can potentially predict the correct semantics by learning from consistent error patterns.

In a second approach, the SLU component is trained on the source data. At decoding time, the input utterance is translated into the source language before being decoded. This approach is referred to as *TrainOnSource*. While this method might be

more sensitive to SMT errors, the SLU model is likely to be more accurate as it is trained on human-annotated data in the source domain.

A third approach is to train models on both the source and target language (*TrainOnSourceAndTarget*), and decode semantics from both the target language utterance and its automatic translation into the source language. This method aims at improving SLU accuracy by capturing patterns that are present in one language but not in the other.

Finally, as *TrainOnTarget* and *TrainOnSource* methods produce different types of errors, a fourth approach is to use each method for different aspects of the SLU process (e.g. predicting the communicative goal of the utterance or the slot/value pairs).

Since cross-lingual SLU depends critically on the quality of both the SLU and SMT components, we have adopted state-of-the-art techniques for both modules in our experiments.

2.1. Semantic tuple classifiers

Discriminative Semantic Tuple Classifiers were recently introduced as an efficient technique for learning to predict semantics from a training set in which each utterance is associated with an utterance-level semantic annotation, but without any word-level semantic alignment [6]. As in previous work, the semantics of the user utterance is a dialogue act representing the user's communicative goal. This dialogue act can be mapped to a tree in which the root node corresponds to the dialogue act type—e.g. informing the system about new constraints (`INFORM`), or rejecting the system's suggestion (`DENY`)—and the branches correspond to slot/value pairs (e.g., `FOOD` → `CHINESE`), as defined in the CUED dialogue act scheme [11].

While each utterance is associated with a semantic representation (e.g., "I would like a Chinese restaurant" is mapped to the tree `INFORM(TYPE(RESTAURANT), FOOD(CHINESE))`), each possible semantic representation cannot be treated as a class because there are not enough examples of individual semantic trees in the data. The STC approach alleviates data sparsity issues by splitting semantic trees into semantic *tuples*—i.e., a sequence of contiguous nodes within a branch of the tree, such as `INFORM` → `TYPE` and `TYPE` → `RESTAURANT`—and then training individual classifiers to predict each tuple from utterance features. When decoding a new utterance, each classifier is applied to the utterance's feature representation, and the semantic tree is reconstructed from the set of predicted tuples (see Figure 1).

The utterance features used to discriminate between semantic concepts consist of the word n-gram counts in the utterance. Each STC is a support vector machine classifier using a linear kernel, and the n-gram size n is optimised for each STC on the training data. Previous work has shown that predicting tuple branches independently of the root yields the best performance on the tourist information domain (*high-recall* approach in [6]), we therefore use the same method for the cross-lingual experiments in this paper (see Figure 1).

The original STC model was extended to output an n-best list of dialogue acts. This is achieved by training probabilistic SVM classifiers that map the classification margin of a tuple t given an user utterance u to a probabilistic confidence score $P(t|u)$. The probability of a 2-level dialogue act da given the user utterance is evaluated by multiplying the probability of the root (da type) with the probability of occurrence of each tuple in the act, together with the complement of the probability of

<s> I want a Chinese restaurant near the NEAR_NAME(railway station) </s>

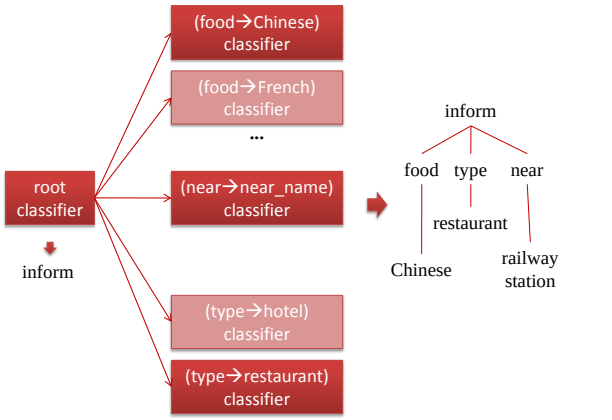

Figure 1: Semantic tree derivation for an utterance in the tourist information domain, with positive concept tuple classifications in darker boxes.

non-occurring tuples:

$$P(da|u) = P(root|u). \prod_{t \in da} P(t|u). \prod_{t \notin da} 1 - P(t|u)$$

Since evaluating this expression for every possible dialogue act would be intractable, the probabilistic STC model returns this probability for every dialogue act containing only tuples with a probability above a cut-off threshold (e.g., $P(t|u) > 0.01$). The function mapping support vector margins to probabilities can then be adjusted to control the precision/recall trade off over predicted slot/value pairs so as to optimise the F-measure on the training set. The results in Section 4 show that this optimisation (referred to as STC++) increases SLU performance, we therefore use it in all our cross-lingual experiments.

2.2. Domain-specific SMT

During the last decade several open-source stochastic machine translation toolkits have been developed (e.g., Moses [7] or Joshua [8]). State-of-the-art SMT techniques use phrase-based models, which learn translation tables mapping bilingual phrase pairs with an associated probability from a corpus of *aligned* sentence pairs. The alignment is typically performed as a preprocessing step using the GIZA++ toolkit [**?**]. Once the translation tables are trained, the optimal translation is found by conducting a beam search over all sequences of translated phrases.

The performance of SMT systems depends on the availability of a parallel data corpus of sufficient size to train the language pair of interest and with linguistic characteristics that match the target task domain. To bootstrap an SMT system, a minimal training set of target data is required. Since our goal is to minimise the required amount of data, we use a *dubbing* approach in which a subset of the source training data is translated by human annotators in order to (1) have an initial dataset to train an SLU system in the target language (re-use of source data avoids the need for a new annotation phase), (2) train a domain-specific SMT system. The trained system can then be used to automatically generate annotated utterances in the target language given the source training set.

In order to minimise the amount of domain-specific data in the target language, we rely on various external resources. First, we use the Europarl corpus [9], which contains over 20

million words in each of the eleven official languages of the European Union, covering the proceedings of the European Parliament 1996-2001. This corpus provides 50k English-French aligned sentence pairs which we use in our experiments. We also use the parallel film subtitle section of the OPUS corpus [10], which contains 240k aligned utterances, and which are closer to the dialogue speaking style than Europarl. These data sets are used together with the translated subset of our domain-specific data (2000 utterances) to train an SMT system using the Moses toolkit.

Moses provides a state-of-the-art system for training a domain specific narrow coverage translation system. To provide a contrast with this approach, we have also evaluated the use of the wide coverage Google Translate tool which although it lacks the specificity of a domain oriented system, it does benefit from being trained on billions of words of both data in the target language and aligned sentence pairs.[1]

3. Experimental method

The remaining of this paper describes our experiments in which the SLU and SMT components detailed in Section 2 are combined to bootstrap a French SLU module from a large set of semantically annotated English utterances (i.e., source language = English, target language = French).

3.1. Data collection

Our domain consists of tourist information dialogues in a fictitious town. The dialogues were collected through user trials in which native speakers of English searched for information about a specific venue by interacting with a dialogue system in a noisy background. These dialogues were previously used for training dialogue management strategies [11]. We use the same training and test sets as in [6] to compare performance of STCs across languages, i.e. 8396 and 1023 transcribed user utterance/dialogue act pairs, respectively, as detailed in Section 2.1. Furthermore, a subset of 406 dialogues have been manually translated into French (including the full test set), yielding 1970 English-French utterance pairs (2k) used for training and 1023 pairs for testing. The utterances were translated by 12 untrained native French speakers from the authors' institutions, taking approximately 15 man-hours, including the collection of French audio recordings to be used for training an automated speech recognition system.

3.2. Cross-lingual SLU systems

The source and target language training sets are combined together to evaluate the cross-lingual SLU methods presented in Section 2, resulting in the following experimental configurations:

1. English STC baseline tested on the English test set as reported in [6].
2. *TrainOnTarget*: STCs trained on 2k manually translated utterances (Man), 8k automatically translated utterances (Auto), or the handcrafted translations together with an automated translation of the rest of the training set (Man+Auto). Automatic translations are obtained either by Google MT or by Moses MT trained on Man only or on Man and external corpora.
3. *TrainOnSource*: English STC baseline tested on automatic translations from French to English.

[1] from Google Translate help section, 01/05/2010.

Conditions	DA	Prec	Rec	F
1. Source language baseline (English)				
STC [6]	94.72	97.19	92.30	94.68
STC++	**95.01**	95.80	95.65	**95.73**
STC++ (trained on 2k)	90.91	92.88	85.58	89.08
2. TrainOnTarget				
Manual (2k)	87.00	83.67	84.44	84.05
Man+Auto (Moses on Man)	90.81	90.13	90.54	90.33
Man+Auto (M. on Man+corpora)[a]	**91.40**	90.17	90.24.	90.20
Auto (Google MT)	90.81	90.77	87.72	89.22
3. TrainOnSource				
Auto (Google MT)	86.71	93.64	77.50	84.81
Auto (Moses on Man+corpora)[b]	90.71	93.83	91.69	**92.75**
4. TrainOnSourceAndTarget				
Bilingual (Man Fr+Moses Auto En)	90.52	94.22	90.69	92.42
Oracle (Man Fr and En tests)	94.43	95.68	94.58	95.13
5. Combining TrainOnTarget and TrainOnSource				
System *a* for root and *b* for slot/values	**91.40**	93.91	91.69	**92.78**

Table 1: Dialogue act type classification accuracy (*DA*), slot/value precision (*Prec*), recall (*Rec*) and F-measure on the test dataset. *Man* = manual translation, *Auto* = automatic translation, *corpora*=general domain bilingual corpora. STC++ is the extended version of STC described in section 2.1.

4. *TrainOnSourceAndTarget*: *bilingual* models trained on concatenations of the English utterances and their translations, by combining the n-gram features of both utterances. At run-time, the system is tested on the French utterances together with their automated English translation (Man Fr+Auto En). As a comparison point, we also evaluate a system using the manual English transcriptions (*Oracle*).
5. Combining *TrainOnSource* and *TrainOnTarget*: the *TrainOnTarget* configuration is used to predict the dialogue act's root node and *TrainOnSource* configuration to predict individual tuples (i.e., slot/value pairs). Since the slot/value pairs included in each dialogue act depend on the predicted root [6], they are likely to differ from the *TrainOnSource* configuration.

4. Evaluation results

Cross-lingual SLU performance is measured in terms of the percentage of correctly classified dialogue act types, as well as the F-measure of the slot/value pairs. Both the slot and the value must be correct to count as a correct classification. The dialogue act type is the root of the output tree, whereas slot/value pairs are trivially extracted from the branches.

The results in Table 1 show that augmenting the target language training set with automatic translations of source language utterances is beneficial (*TrainOnTarget*), since the slot/value F-measure increases from .84 to .90 when adding 8k automatic translations to the 2k manually translated utterances. Similarly, the dialogue act type classification accuracy increases from 87% to 90.8% when training Moses on the 2k in-domain utterance pairs, and up to 91.4% when adding the out-of-domain bilingual corpora detailed in Section 2.2. Additionally, we find that Moses offers only a slight performance increase over Google Translate, which is consistent with their comparable BLEU scores (see Table 2). This suggests that general-domain SMT systems can be used for cross-lingual SLU to further re-

SMT System	BLEU Score
English → French	
Google Translate	51
Moses MT (2k Manual data)	46
Moses MT (2k Manual data and corpora)	50
French → English	
Google Translate	58
Moses MT (2k Manual data and corpora)	61

Table 2: BLEU scores [**?**] over the test set translated using Google Translate and Moses. The French test set is the manual translation of the English test set (1023 utterances).

duce development cost.

While cross-lingual SLU models do not perform as well as models trained and tested on the source language (STC++ baseline on English), the performance decrease observed between an English and French STC model trained on the same 2k utterances (F=.89 and .84 respectively) suggests that this difference is due to inherent linguistic differences between the two languages (e.g., lexical variability) rather than the quality of the training data.

Interestingly, results for the *TrainOnSource* configuration in Table 1 show that using Moses to translate the input utterance into the source language before decoding yields the best performance in terms of F-measure ($F = .93$), while the dialogue act classification accuracy remains slightly worse than with the *TrainOnTarget* systems. This suggests that both methods should be used complementarity, as confirmed by the experimental results presented at the bottom of Table 1 which lead to the best overall performance.

Additionally, the gap between the SMT systems is larger when using the *TrainOnSource* approach ($F = .85$ vs. $F = .93$). This is likely to be due to the fact that the SLU module trained on the source language cannot learn to correct error patterns introduced by the SMT system, hence the differences in SMT quality from French to English reported in Table 2) have a larger impact on the SLU accuracy.

Finally, we find that combining both source and target language data at training and decoding time does not improve performance. Furthermore, combining the French utterances together with an oracle English translation does not improve performance over the baseline English STC++ system, suggesting that the French translations do not provide additional information over the English utterances.

5. Conclusion

This paper has investigated methods for bootstrapping an SLU component in a new target language from an existing set of semantically-annotated utterances in a source language. Since word-level semantic annotations are costly, Semantic Tuple Classifiers (STCs) are used in conjunction with statistical machine translation models both of which are trained from *unaligned* data to further reduce development time. Results show that training STCs on automatically translated data produced the best performance for predicting the utterance's dialogue act type, however individual slot/value pairs are best predicted by training STCs on the source language and using them to decode translated utterances. Overall, results show that good performance can be achieved by training the SMT system on a relatively small parallel corpus.

In addition to this work, a French speech recognition module was trained on audio recordings of the translated utterances, thus making it possible to port a whole dialogue system from English to French.[2] The resulting system can now be used to collect French conversational data in more realistic noisy environments to refine both the speech recognition and SLU components, and to validate the results presented in this paper.

6. References

[1] E. Levin and R. Pieraccini, "Chronus, the next generation," in *Proceedings of the ARPA Workshop on Spoken Language Technology*, 1995.

[2] R. Schwartz, S. Miller, D. Stallard, and J. Makhoul, "Language understanding using hidden understanding models," in *Proceedings of ICSLP*, 1996.

[3] Y. He and S. Young, "Spoken language understanding using the hidden vector state model," *Speech Communication*, vol. 48, no. 3-4, pp. 262–275, 2006.

[4] Y.-Y. Wang and A. Acero, "Discriminative models for spoken language understanding," in *Proceedings of ICSLP*, 2006.

[5] F. Mairesse, M. Gašić, F. Jurčíček, S. Keizer, B. Thomson, K. Yu, and S. Young, "Spoken language understanding from unaligned data using discriminative classification models," in *Proceedings of ICASSP*, 2009.

[6] M.-J. Meurs, F. Lefèvre, and R. De Mori, "Spoken language interpretation: On the use of dynamic bayesian networks for semantic composition," in *Proceedings of ICASSP*, 2009.

[7] S. Young, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu, "The hidden information state model: A practical framework for POMDP-based spoken dialogue management," *Computer Speech & Language*, vol. 24, no. 2, pp. 150–174, April 2010.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of ACL*, ser. demo session, 2007.

[9] Z. Li, C. Callison-Burch, C. Dyer, S. Khudanpur, L. Schwartz, W. Thornton, J. Weese, and O. Zaidan, "Joshua: An open source toolkit for parsing-based machine translation," in *Proceedings of Statistical Machine Translation*, 2009.

[10] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT Summit*, 2005.

[11] J. Tiedemann, "News from OPUS - A collection of multilingual parallel corpora with tools and interfaces," in *Recent Advances in Natural Language Processing*, vol. V, 2009, pp. 237–248.

[2] The system prompts were also manually translated.