Cross-lingual projection for class-based language models

Beat Gfeller and **Vlad Schogol** and **Keith Hall**
Google Inc.
{beatg,vlads,kbhall}@google.com

Abstract

This paper presents a cross-lingual projection technique for training class-based language models. We borrow from previous success in projecting POS tags and NER mentions to that of a trained class-based language model. We use a CRF to train a model to predict when a sequence of words is a member of a given class and use this to label our language model training data. We show that we can successfully project the contextual cues for these classes across pairs of languages and retain a high quality class model in languages with no supervised class data. We present empirical results that show the quality of the projected models as well as their effect on the down-stream speech recognition objective. We are able to achieve over 70% of the WER reduction when using the projected class models as compared to models trained on human annotations.

1 Introduction

Class-based language modeling has a long history of being used to improve the quality of speech recognition systems (Brown et al., 1992; Knesser and Ney, 1993). Recent work on class-based models has exploited named entity recognition (NER) approaches to label language model training data with class labels (Levit et al., 2014; Vasserman et al., 2015), providing a means to assign words and phrases to classes based on their context. These contextually assigned classes have been shown to improve speech recognition significantly over grammar-based, deterministic class assignments.

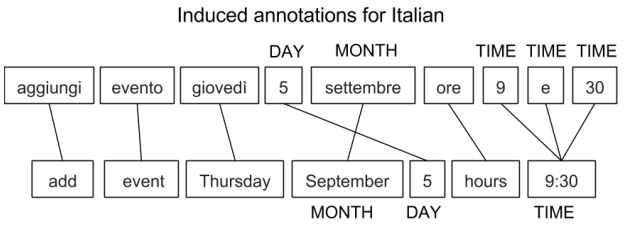
In this work, we address the problem of labeling training data in order to build a class sequence tagger. We borrow from the successes of previous cross-lingual projection experiments for labeling tasks (Yarowsky et al., 2001; Yarowsky and Ngai, 2001; Burkett et al., 2010; Padó and Lapata, 2009). We focus on *numeric* classes (e.g., address numbers, dates, currencies, times, etc.) as the sequence-based labeling approach has been shown to be effective for identifying them. Given a model trained from human-labeled data in one language (we refer to this as the high-resource language), we label translations of sentences from another language (referred to as the low-resource language). We show that we can project the numeric entity boundaries and labels across the aligned translations with a phrase-based translation model. Furthermore, we show that if we train a class labeling model on the projected low-resource language and then use that to build a class-based speech recognition system, we achieve between 70% and 85% of the error reduction as we would have achieved with human-labeled examples in the low-resource language.

We present empirical results projecting numeric entity labels from English to Russian, Indonesian, and Italian. We present full speech recognition results for using human annotated data (the ideal performance) and projected data with various sizes of training data.

2 Related work

There is an increasingly large body of work based on exploiting alignments between translations of sentences in multiple languages (Yarowsky et al., 2001; Yarowsky and Ngai, 2001; Burkett et al., 2010; Das and Petrov, 2011). In this work we employ the simple approach of projecting annotations across alignments of translated sentences. Our cross-lingual approach is closely related to other NER projection approaches (Huang et al.,

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 83–88,
Berlin, Germany, August 7-12, 2016.



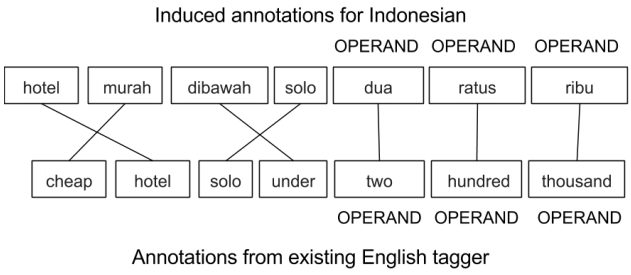


Figure 1: Examples of cross-lingual projection for numeric entities.

2003; Moore, 2003); however, we have focused on a limited class of entities which may explain why the simple approach works reasonably well.

Our projection approach is most closely related to that presented in (Yarowsky et al., 2001) and (Padó and Lapata, 2009). In each of these, labels over sequences of words are projected across alignments directly from one language to the other. While we follow a similar approach, our goal is not necessarily to get the exact projection, but to get a projection which allows us to learn contextual cues for the classes we are labeling. Additionally, we focus on the case where we are generating the translated data rather that identifying existing parallel data. Similar to (Yarowsky and Ngai, 2001), we filter out poor alignments (details are described in Section 3.2).

3 Methodology

3.1 Training class taggers for language modeling

We use a statistical sequence tagger to identify and replace class instances in raw text with their label. For example, the tokens *10 thousand dollars* in the raw training text may be replaced with a placeholder class symbol. The decision is context-dependent: the tagger is able to resolve ambiguities among possible labels, or even leave the text unchanged. Next, this modified text is used to train a standard n-gram language model. Finally, all placeholders become non-terminals in the language model and are expanded either statically or dynamically with stochastic finite-state class grammars (see Figure 2 for an example). Decorator tokens inside the grammars are used to mark class instances in the word lattice so that they can be converted (after recognition) to the desired written forms using deterministic spoken-to-written text-normalization rules.

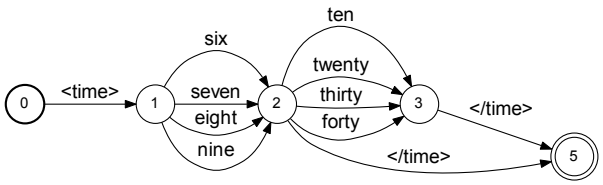


Figure 2: This FST is a small excerpt of the full grammar for *TIME*. Arc weights are not shown.

3.2 Cross-lingual Projection Techniques

The starting point for cross-lingual projection is to train a statistical sentence tagger of high quality in a high-resource language, i.e., a language where both a lot of training data and human annotators are readily available. We use English in our experiments.

To obtain annotated sentences in a low-resource language, we translate unlabeled sentences into the high-resource language. We use an in-house phrase-based statistical machine translation system (Koehn et al., 2003) which is trained with parallel texts extracted from web pages; described in detail in Section 4.1 of (Nakagawa, 2015). The translation system we use provides token-by-token alignments as part of the output. This is achieved by keeping alignments along with phrase-pairs during the phrase extraction stage of training the alignment system.

The high quality sentence tagger is applied to the translated sentences. Then, using the alignments between the translated sentences, we map class tags back to the low-resource language. See Figure 1 for examples of actual mappings produced by this procedure.

With this approach, we can produce arbitrarily large in-domain annotated training sets for the low-resource language. These annotated sentences are then used to train a class tagger for the low-resource language. The main question is whether the resulting class tagger is of sufficient quality for our down-stream objective.

For the goal of training a class-based language model in a low-resource language, one may consider a different approach than the one just described: instead of training a tagger in the low-resource language, each sentence in the language model training data could be translated to the high-resource language, tagged using the statistical tagger, and projected back to the low-resource language. The primary reason for not pursuing this approach is the size of the language model training data (tens of billions of sentences). Translating a corpus this large is prohibitive. As the high-resource language tagger is trained on approximately 150K tokens, we believe that we have covered a large number of the predictive cues for the set of classes.

Alignment details

When projecting the class labels back from a translated sentence to the original sentence, various subtle issues arise. We describe these and our solutions for each in this section.

To tag a token in the low-resource language, we see which tokens in the high-resource language are aligned to it in the translation, and look at their class tags. If all of these tokens have the same class tag, we assign the same tag to the low-resource language token. Otherwise, we use the following rules:

- If some tokens have no class tag but others have some class tag, we still assign the class tag to the original token.

- If multiple tokens with different class tags map to the original token, we consider the tagging ambiguous. In such a case, we simply skip the sentence and do not use it for training the low-resource tagger. We can afford to do so because there is no shortage of unlabeled training sentences.

In a number of cases, we ignore sentence pairs which may have contained alignments allowing us to project labels, but also contained noise (e.g., spurious many-to-one alignments). We rejected poor alignments 2%, 31% and 14% of the time for Indonesian, Russian and Italian respectively. Date and time expressions were often affected by these noisy alignments.

4 Empirical evaluation

4.1 Data

We trained an English conditional random field (CRF) (Lafferty et al., 2001) tagger to be used in all experiments in order to provide labels for the sentences produced by translation. To train this tagger we obtained a data set of 24,503 manually labeled sentences (150K tokens) sampled from a corpus of British English language model training material. Each token is labeled with one of 17 possible tags. About 95% of the tokens are labeled with a 'none' tag, meaning that the token is not in any of the pre-determined non-lexical classes.

Separately, we obtained similar training sets to create Italian, Indonesian and Russian taggers. The models trained from these labeled data sets were used only to create baseline systems for comparison with the cross-lingual systems.

To provide input into our cross-lingual projection procedure, we also sampled datasets of unlabeled sentences of varying sizes for each evaluation language, using the same sampling procedure as used for the human-labeled sets.

Note that these tagger training sets have inconsistent sizes across languages (see Table 2) due to the nature of the sampling procedure: Each training source is searched for sentences matching an extensive list of patterns of numeric entities. Sentences from each training source are collected up to a source-specific maximum number (which may not always be reached). We also apply a flattening step to increase diversity of the sample.

4.2 CRF model

Our CRF tagger model was trained online using a variant of the MIRA algorithm (Crammer and Singer, 2003). Our feature set includes isolated features (for word identity w_i, word type d_i, and word cluster c_i) as well as features for neighboring words $w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}, w_{i+3}$, neighboring clusters $c_{i-2}, c_{i-1}, c_{i+1}, c_{i+2}, c_{i+3}$, pair features $(w_i, d_{i-1}), (w_i, d_{i+1}), (d_i, d_{i-1}), (d_i, d_{i+1})$, and domain-specific features (indicators for tokens within a given numeric range, or tokens that end in a certain number of zero digits). We also include class bias features, which capture the class prior distribution found in the training set.

4.3 Metrics

We use two manually transcribed test sets to evaluate the performance of our approach in the con-

Test Set	Utts	Words	% Numeric words
NUM ID	9,744	60,781	19%
NUM RU	10,988	59,933	22%
NUM IT	8,685	48,195	18%
VS ID	9,841	36,276	2%
VS RU	12,467	49,403	3%
VS IT	12,625	47,867	2%

Table 1: *NUM* refers to the NUMERIC entities test set and *VS* refers to the VOICE-SEARCH test set.

text of numeric transcription. The first test set VOICE-SEARCH (approximately 48K words for Italian and Russian, and approximately 36K words for Indonesian) is a sample from general voice-search traffic, and tracks any regressions that appear as a result of biasing too heavily toward the selected classes. The other test set NUMERIC (approximately 48K words for Italian, and approximately 60K for Russian and Indonesian) contains utterances we expect to benefit from class-based modeling of numeric entities. See Table 1 for details on these test sets.

We report word-error-rate (WER) on each test set for each model evaluated, including two baseline systems (one built without classes at all and another that has classes identified by a tagger trained on human-labeled data). We also report a labeled-bracket F1 score to show the performance of the tagger independent of the speech-recognition task. For each language, the test set used for labeled-bracket F1 is a human-labeled corpus of approximately 2K sentences that were held out from the human-labeled corpora for the baseline systems.

4.4 Results

The results in Table 2 show that all class-based systems outperform the baseline in WER on the NUMERIC test set, while performance on the VOICE-SEARCH test set was mostly flat. The flat performance on VOICE-SEARCH is expected: as seen in Table 1 this test set has a very low proportion of words that are numeric in form. We provide results on this test set in order to confirm that our approach does not harm general voice-search queries. As for performance on the NUMERIC test set, larger cross-lingual data sets led to better performance for Russian and Italian, but caused a slight regression for Indonesian. The translation system we use for these experiments has been optimized for a general-purpose web search

Model	F1	*NUM* WER	*VS* WER
ID Baseline (no classes)	-	20.0	10.1
ID Cross-lingual 15K	0.64	19.3	10.1
ID Cross-lingual 37K	0.65	19.4	10.1
ID Cross-lingual 77K	0.64	19.5	10.1
ID Human-labeled	0.83	19.1	10.1
RU Baseline (no classes)	-	28.7	17.1
RU Cross-lingual 16K	0.37	26.4	17.0
RU Cross-lingual 98K	0.39	26.2	17.1
RU Human-labeled	0.87	25.3	16.8
IT Baseline (no classes)	-	23.0	14.8
IT Cross-lingual 18K	0.55	19.7	14.8
IT Cross-lingual 104K	0.57	19.6	14.8
IT Human-labeled	0.88	19.0	14.8

Table 2: *NUM* refers to the NUMERIC entities test set and *VS* refers to the VOICE-SEARCH test set. All *NUM* WER results are statistically significant ($p < 0.1\%$) using a paired random permutation significance test.

translation task rather than for an academic task. When evaluated on a test set matched to the translation task, performance for Russian-to-English was considerably worse than for Indonesian-to-English or Italian-to-English.

For Indonesian (ID), the human-labeled system achieved a 4.5% relative WER reduction on NUMERIC, while the best cross-lingual system achieved a 3.5% relative reduction.

For Russian (RU), the human-labeled system improved more, achieving an 11.8% relative reduction on NUMERIC, while the best cross-lingual system achieved an 8.7% relative reduction.

Finally, for Italian (IT), the human-labeled system gave an impressive 17.4% relative reduction on NUMERIC, while the best cross-lingual system achieved a 14.8% relative reduction on the same test set.

Across the three languages, the cross-lingual systems achieved relative error reductions on the NUMERIC test set that were between 70% and 85% of the reduction achieved when using only human-labeled data for training the class tagger.

4.5 Error Analysis

We noticed that the Russian cross-lingual-derived training set was of lower quality than those of the other languages, as seen in the labeled-bracket F1 metric in Table 2. Looking more closely, we

noticed that the per-class F1 scores tended to be lower for labels used for dates and times. This observation also concides with the observation that the alignment procedure frequently ran into ambiguity issues when aligning month, day and year tokens between Russian and English, thus significantly reducing the coverage of these labels in the induced cross-lingual training set.

5 Conclusion

We presented a cross-lingual projection technique for training class-based language models. We extend a previously successful sequence-modeling-based class labeling approach for identifying contextually-dependent class assignments by projecting labels from a high-resource language to a low-resources language. This allows us to build class-based language models in low-resource languages with no annotated data. Our empirical results show that we are able to achieve between 70% and 85% of the error reduction that we would have obtained had we used human-labeled data.

While cross-lingual projection for sequence-labeling techniques are well known in the community, our approach exploits the fact that we are generating training data from the projection rather than using the projected result directly. Furthermore, noise in the class-labeling system does not cripple the language model as it learns a distribution over labels (including no label).

In future work, we will experiment with alternative projection approaches including projecting the training data and translating from the high-resource language to the low-resource language. We also plan to experiment with different projection approaches to address the ambiguity issues we observed when aligning time and date expressions.

6 Acknowledgments

We would like to thank the anonymous reviewers for their detailed reviews and suggestions.

References

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

David Burkett, Slav Petrov, John Blitzer, and Dan Klein. 2010. Learning better monolingual models with unannotated bilingual text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 46–54. Association for Computational Linguistics.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *J. Mach. Learn. Res.*, 3:951–991, March.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 600–609. Association for Computational Linguistics.

Fei Huang, Stephan Vogel, and Alex Waibel. 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition - Volume 15*, MultiNER '03, pages 9–16. Association for Computational Linguistics.

Reinhard Knesser and Hermann Ney. 1993. Improved clustering techniques for class-based statistical language modelling. In *Proc. Eurospeech*. ISCA - International Speech Communication Association, September.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1 of *NAACL '03*, pages 48–54. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Michael Levit, Sarangarajan Parthasarathy, Shuangyu Chang, Andreas Stolcke, and Benoit Dumoulin. 2014. Word-phrase-entity language models: Getting more mileage out of n-grams. In *Proc. Interspeech*. ISCA - International Speech Communication Association, September.

Robert C. Moore. 2003. Learning translations of named-entity phrases from parallel corpora. In *Proceedings of the Tenth Conference on European*

Chapter of the Association for Computational Linguistics - Volume 1, EACL '03, pages 259–266. Association for Computational Linguistics.

Tetsuji Nakagawa. 2015. Efficient top-down BTG parsing for machine translation preordering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL 2015, Volume 1: Long Papers*, pages 208–218.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340, September.

Lucy Vasserman, Vlad Schogol, and Keith Hall. 2015. Sequence-based class tagging for robust transcription in asr. In *Proc. Interspeech*. ISCA - International Speech Communication Association, September.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8. Association for Computational Linguistics.