Learning Bilingual Categories in Unsupervised Inversion Transduction Grammar Induction

Markus SAERS and **Dekai WU**
Human Language Technology Center
Dept. of Computer Science and Engineering
Hong Kong University of Science and Technology
{masaers|dekai}@cs.ust.hk

Abstract

We present the first known experiments incorporating unsupervised bilingual nonterminal category learning within end-to-end fully unsupervised transduction grammar induction using matched training and testing models. Despite steady recent progress, such induction experiments until now have not allowed for learning differentiated nonterminal categories. We divide the learning into two stages: (1) a *bootstrap* stage that generates a large set of categorized short transduction rule hypotheses, and (2) a *minimum conditional description length* stage that simultaneously prunes away less useful short rule hypotheses, while also iteratively segmenting full sentence pairs into useful longer categorized transduction rules. We show that the second stage works better when the rule hypotheses have categories than when they do not, and that the proposed conditional description length approach combines the rules hypothesized by the two stages better than a mixture model does. We also show that the compact model learned during the second stage can be further improved by combining the result of different iterations in a mixture model. In total, we see a jump in BLEU score, from 17.53 for a standalone minimum description length baseline with no category learning, to 20.93 when incorporating category induction on a Chinese–English translation task.

1 Introduction

Even simple lexical translations are surprisingly context-dependent, in this paper we aim to learn a translation model that can base contextual translation decision on more than lexical n-grams, both in the input and output language. In a syntactic translation system such as inversion transduction grammars (ITGs), this can be achieved with unsupervised bilingual category induction. Surface-based and hierarchical models only use output language n-grams, and syntactic model typically choose the categories from either the input or the output language, or attempts to heuristically synthesize a set of bilingual categories from the two monolingual sets. In contrast, we attempt to learn a set of bilingual categories without supervision, which gives a unique opportunity to strike a good balance between the two approaches.

The specific translation of words and segments depend heavily on the context. A grammar-based translation model can model the context with nonterminal categories, which allows (a) moving beyond n-grams (as a compliment to the language model prior which is typically preserved), and (b) taking both the input and output language context into account. Typical syntactic MT systems either ignore categories (bracketing ITGs and hierarchical models), or derive the categories from tree-banks, which relies on choosing the set of categories from either language, or heuristically synthesize it from both; both approaches eliminates the full benefits of (b). In contrast, unsupervised induction of a bilingual category set has the potential to fully take advantage of (b).

Recent work has seen steady improvement in translation quality for completely unsupervised transduction grammar induction under end-to-end purely matched training and testing model conditions. In this paper, we take a further step along this line of research by incorporating unsupervised bilingual category induction into the learning process. To our knowledge, no previous attempt has been made to incorporate bilingual categories under such conditions. Matching the training and testing models as closely as possible is

a fundamental principle taken for granted in most applications of machine learning, but for machine translation it has been the norm to see very different assumptions during training and testing, which makes it difficult to assess the effects of changing or tweaking the model—the observed effect may not be repeatable. By matching training and testing conditions, this risk is minimized.

A bilingual category is similar to a monolingual category in that it is realized as the left-hand side label of a (transduction) grammar rule, but differ in what it represents. A monolingual category only encodes how something relates to other parts of the language, a bilingual category should encode how a translation equivalence relates to other translation equivalences. It needs to account for the relationship between two languages as well as the relationship between the parts of the individual languages. This makes the usage of existing tagging schemes problematic. It would be possible to use the categories from either of the languages (assuming they are languages with enough resources) and impose these on the other language. This could work for closely related languages, but we are translating between English and Chinese: two very different languages, and we know that the category sets of either language is a poor fit for the other. Another possibility is to take the cross-product of the monolingual category sets, but handling such a large set of categories becomes unwieldy in ITG induction, a process which is resource intensive as is, without exploding the set of nonterminals. Instead, we opt for unsupervised learning of the bilingual categories during induction of the ITG itself.

The novel learning method we propose consists of an initial hypothesis generator that proposes (a) short lexical translations and (b) nonterminal categories, screened by a mechanism that (c) verifies the usefulness of the hypotheses while (d) uses them to further generate longer transduction rules. For convenience, our implementation breaks this into two stages: one that generates a large set of short transduction rule hypotheses, and another that iteratively segments long transduction rules (initialized from the sentence pairs in the training data) by trying to reuse a minimal subset of the hypotheses while chipping away at the long sentence pair rules until the conditional description length is minimized.

The paper is structured so that, after giving the background situated within the context of relevant related research (Section 2), we define the proposed conditional description length approach, which represents the ideal model search (Section 3). We then detail the two stages of our proposed learning algorithm, which represents our approximation of the search problem (Sections 4 and 5). After the theory we detail the particular experiments we conducted (Section 6) and the results from those experiments (Section 7). Finally, we offer some conclusions (Section 8).

2 Background

Description length has been used before to drive iterative segmenting ITG learning (Saers et al., 2013). We will use their algorithm as our baseline, but the simple mixture model we used then works poorly with our ITG with categories. Instead, we propose a tighter incorporation, where the rule segmenting learning is biased towards rules that are present in the categorized ITG.

We refer to this objective as minimizing *conditional description length*, since technically, the length of the ITG being segmented is *conditioned* on the categorized ITG. Conditional description length (CDL) is detailed in Section 3. The minimum CDL (MCDL) objective differs from the simple mixture model in that it separates the rule hypotheses into two groups: the ones that are used during segmentation and therefor carries over to the final induced ITG, and those that do not and are effectively filtered out. As we will see, MCDL far outperforms the mixture model when one of the ITGs has categories and the other does not.

A problem with the description length family of learning objectives is that they tend to commit to a segmentation when it would be wise to keep the unsegmented rule *as well*—a significant part of the success of phrase-based translation models comes from their approach to keep all possible segmental translations (that do not violate the prerequisite word alignment). We will show that we can counter this by combining different iterations of the same segmentation process into a single grammar, which gives a significant bump in BLEU scores.

By insisting on the fundamental machine learning principle of matching the training model to the testing model, we do forfeit the short term boost in BLEU that is typically seen when embedding a learned ITG

in the midst of the common heuristics employed in statistical machine translation. For example, Cherry and Lin (2007), Zhang et al. (2008), Blunsom et al. (2008), Blunsom et al. (2009), Haghighi et al. (2009), Saers and Wu (2009), Blunsom and Cohn (2010), Burkett et al. (2010), Riesa and Marcu (2010), Saers et al. (2010), Saers and Wu (2011), Neubig et al. (2011), and Neubig et al. (2012) all plug some aspect of the ITGs they learn into training pipelines for existing, mismatched decoders, typically in the form of the word alignment that an ITG imposes on a parallel corpus as it is biparsed. Our own past work has also taken similar approaches, but it is not necessary to do so—instead, any ITG can be used for decoding by directly parsing with the input sentence as a hard constraint, as we do in this paper. Although it allows you to tap into the vast engineering efforts that have gone into perfecting existing decoders, it also prevents you from surpassing them in the long run. The motivation for our present series of experiments is that, as a field we are well served by tackling the fundamental questions as well, and not exclusively focusing on engineering short term incremental BLEU score boosts where the quality of an induced ITG itself is obscured because it is embedded within many other heuristic algorithms.

When the structure of an ITG is induced without supervision, it is possible to get an effect that resembles MDL. Zhang et al. (2008) impose a sparsity prior over the rule probabilities to prevent the search from having to consider all the rules found in the Viterbi biparses. Blunsom et al. (2008), Blunsom et al. (2009), Blunsom and Cohn (2010), Neubig et al. (2011), and Neubig et al. (2012) use Gibbs sampling to learn ITGs with priors over the rule structures that serve a similar purpose to the model length component of description length. All of the above evaluate their models by biparsing the training data and feeding the imposed word alignment into an existing, mismatched SMT learning pipeline.

Transduction grammars can also be induced with supervision from treebanks, which cuts down the search space by enforcing external constraints (Galley et al., 2006). Although this constitutes a way to borrow nonterminal categories that help the translation model, it complicates the learning process by adding external constraints that are bound to match the translation model poorly.

3 Conditional description length

Conditional description length (CDL) is a general method for evaluating a model and a dataset given a preexisting model. This makes it ideal for augmenting an existing model with a variant model of the same family. In this paper we will apply this to augment an existing inversion transduction grammar (ITG) with rules that are found with a different search strategy. CDL is similar to description length (Solomonoff, 1959; Rissanen, 1983), but the length calculations are subject to additional constraints. When minimum CDL (MCDL) is used as a learning objective, all the desired properties of minimum description length (MDL) are retained: the model is allowed to become less certain about the data provided that the it shrinks sufficiently to compensate for the loss in precision. MDL is a good way to prevent over-fitting, and MCDL retains this property, but for the task of inducing a model that is specifically tailored toward augmenting an existing model. Formally, the conditional description length is:

$$DL\left(\Phi, D|\Psi\right) = DL\left(D|\Phi, \Psi\right) + DL\left(\Phi|\Psi\right)$$

where Ψ is the fixed preexisting model, Φ is the model being induced, and D is the data. The total unconditional length is:

$$\begin{aligned} &DL\left(\Psi, \Phi, D\right) = \\ &\qquad DL\left(D|\Phi, \Psi\right) + DL\left(\Phi|\Psi\right) + DL\left(\Psi\right) \end{aligned}$$

In minimizing CDL, we fix $DL\left(\Psi\right)$ instead of allowing Ψ to vary as we would in full MCDL; to be precise, we seek:

$$\begin{aligned} &\operatorname*{argmin}_{\Phi} DL\left(\Psi, \Phi, D\right) \\ &= \operatorname*{argmin}_{\Phi} DL\left(D|\Phi, \Psi\right) + DL\left(\Phi|\Psi\right) + DL\left(\Psi\right) \\ &= \operatorname*{argmin}_{\Phi} DL\left(\Phi, D|\Psi\right) \\ &= \operatorname*{argmin}_{\Phi} DL\left(D|\Phi, \Psi\right) + DL\left(\Phi|\Psi\right) \end{aligned}$$

To measure the CDL of the data, we turn to information theory to count the number of bits needed to encode the data given the two models under an optimal encoding (Shannon, 1948), which gives:

$$DL\left(D|\Phi, \Psi\right) = -\lg P\left(D|\Phi, \Psi\right)$$

To measure the CDL of the model, we borrow the encoding scheme for description length presented in Saers et al. (2013), and define the conditional description length as:

$$DL\left(\Phi|\Psi\right) \equiv DL\left(\Phi\right) - DL\left(\Phi \cap \Psi\right)$$

To determine whether a model Φ has a shorter conditional description length, than another model Φ', it is sufficient to be able to subtract one length from the other. For the model length, this is trivial as we merely have to calculate the length of the difference between the two models in our theoretical encoding. For data length, we need to solve:

$$\begin{aligned} &DL\left(D|\Phi', \Psi\right) - DL\left(D|\Phi, \Psi\right) \\ &\quad = -\lg P\left(D|\Phi', \Psi\right) - -\lg P\left(D|\Phi, \Psi\right) \\ &\quad = -\lg \frac{P\left(D|\Phi', \Psi\right)}{P\left(D|\Phi, \Psi\right)} \end{aligned}$$

4 Generating rule hypotheses

In the first stage of our learning approach, we generate a large set of possible rules, from which the second stage will choose a small subset to keep. The goal of this stage is to keep the *recall* high with respect to a theoretical "optimal ITG", *precision* is achieved in the second stage. We rely on chunking and category splitting to generate this large set of rule hypotheses.

To generate these high-recall ITGs, we will follow the bootstrapping approach presented in Saers et al. (2012), and start with a finite-state transduction grammar (FSTG), do the chunking and category splitting within the FSTG framework before transferring the resulting grammar to a corresponding ITG. This is likely to produce an ITG that performs poorly on its own, but may be informative in the second stage.

5 Segmenting rules

In the second stage of our learning approach, we segment rules explicitly representing the entire training data, into smaller—more general—rules, reusing rules from the first stage whenever we can. By driving the segmentation-based learning with a minimum description length objective, we are learning a very concise ITG, and by conditioning the description length on the rules hypothesized in the first stage, we separate the good rule hypotheses from the bad: the good rules—along with their categorizing left-hand sides—are reused and the bad are not.

In this work, we are only considering segmentation of lexical rules, which keeps the ITG in normal form, greatly simplifying processing without altering the expressivity. A lexical ITG rule has the form $A \to e_{0..T}/f_{0..V}$, where A is the left-hand side nonterminal—the category, $e_{0..T}$ is a sequence of T (from position 0 up to but not including position T) L_0 tokens and $f_{0..V}$ is a sequence of V (from position 0 up to but not including position V) L_1 tokens. When segmenting this rule, three new rules are produced which take one of the following forms depending on whether the segmentation is inverted or not:

$$\begin{aligned} A &\to [BC] & & A \to \langle BC \rangle \\ B &\to e_{0..S}/f_{0..U} & \text{or} \quad & B \to e_{0..S}/f_{U..V} \\ C &\to e_{S..T}/f_{U..V} & & C \to e_{S..T}/f_{0..U} \end{aligned}$$

All possible splits of the terminal rule can be accounted for by choosing the identities of B, C, S and U, as well as whether the split it straight or inverted.

Algorithm 1 Iterative rule segmenting learning driven by minimum conditional description length.

```
Φ                                  ▷ The ITG being induced
Ψ                ▷ The ITG the learning is conditioned on
repeat
  δ_sum ← 0
  bs ← collect_biaffixes(Φ)
  bδ ← []
  for all b ∈ bs do
    δ ← eval_cdl(b, Ψ, Φ)
    if δ < 0 then
      bδ ← [bδ, ⟨b, δ⟩]
  sort_by_delta(bδ)
  for all ⟨b, δ⟩ ∈ bδ do
    δ' ← eval_cdl(b, Ψ, Φ)
    if δ' < 0 then
      Φ ← make_segmentations(b, Φ)
      δ_sum ← δ_sum + δ'
until δ_sum ≥ 0
return Φ
```

The pseudocode for the iterative rule segmenting learning algorithm driven by minimal conditional description length can be found in Algorithm 1. It uses the methods `collect_biaffixes`,

`eval_cdl`, `sort_by_delta` and `make_segmentations`. These methods collect all biaffixes in the rules of an ITG, evaluate the difference in conditional description length, sorts candidates by these differences, and commits to a given set of candidates, respectively. To evaluate the CDL of a proposed set of candidate segmentations, we need to calculate the difference in CDL between the current model, and the model that would result from committing to the candidate segmentations:

$$\begin{aligned} &DL\left(D, \Phi'|\Psi\right) - DL\left(D, \Phi|\Psi\right) \\ &\quad = DL\left(D|\Phi', \Psi\right) - DL\left(D|\Phi, \Psi\right) \\ &\qquad + DL\left(\Phi'|\Psi\right) - DL\left(\Phi|\Psi\right) \end{aligned}$$

The model lengths are trivial, as we merely have to encode the rules that are removed and inserted according to our encoding scheme and plug in the summed lengths in the above equation. This leaves the length of the data, which would be:

$$DL\left(D|\Phi', \Psi\right) - DL\left(D|\Phi, \Psi\right) = -\lg \frac{P\left(D|\Phi', \Psi\right)}{P\left(D|\Phi, \Psi\right)}$$

For the sake of convenience in efficiently calculating this probability, we make the simplifying assumption that:

$$P\left(D|\Phi, \Psi\right) \approx P\left(D|\Phi\right) = P\left(D|\theta\right)$$

where θ represents the model parameters, which reduces the difference in data CDL to

$$-\lg \frac{P\left(D|\theta'\right)}{P\left(D|\theta\right)}$$

which lets us determine the probability through biparsing with the model being induced. Biparsing is, however, a very expensive operation, and we are making relatively small changes to the ITG, so we will further assume that we can estimate the CDL difference in closed form based on the model parameters. Given that we are splitting the rule r_0 into the three rules r_1, r_2 and r_3, and that the probability mass of r_0 is distributed uniformly over the new rules, the new grammar parameters θ' will be identical to θ, except that:

$$\begin{aligned} \theta'_{r_0} &= 0 \\ \theta'_{r_1} &= \theta_{r_1} + \frac{1}{3}\theta_{r_0} \\ \theta'_{r_2} &= \theta_{r_2} + \frac{1}{3}\theta_{r_0} \\ \theta'_{r_3} &= \theta_{r_3} + \frac{1}{3}\theta_{r_0} \end{aligned}$$

We estimate the CDL of the corpus given this new parameters to be:

$$-\lg \frac{P\left(D|\theta'\right)}{P\left(D|\theta\right)} \approx -\lg \frac{\theta'_{r_1}\theta'_{r_2}\theta'_{r_3}}{\theta_{r_0}}$$

To generalize this to a set of rule segmentations, we construct the new parameters θ' to reflect all the changes in the set in a first pass, and then sum the differences in CDL for all the rule segmentations with the new parameters in a second pass.

6 Experimental setup

The learning approach we chose has two stages, and in this section we describe the different ways of using these two stages to arrive at a final ITG, and how we intend to evaluate the quality of those ITGs.

For the first stage, we will use the technique described in Saers et al. (2012) to start with a finite-state transduction grammar (FSTG) and perform chunking before splitting the nonterminal categories and moving the FSTG into ITG form. We will perform one round of chunking, and two rounds of category splitting (resulting in 4 nonterminals and 4 preterminals, which becomes 8 nonterminals in the ITG form). Splitting all categories is guarnteed to at least double the size of the grammar, which makes is impractical to repeat more times. At each stage, we run a few iterations of expectation maximization using the algorithm detailed in Saers et al. (2009) for biparsing. For comparison we also bootstrap a comparable ITG that has not had the categories split. Before using either of the bootstrapped ITGs, we eliminate all rules that do not have a probability above a threshold that we fixed to 10^{-50}. This eliminates the highly unlikely rules from the ITG.

For the second stage, we use the iterative rule segmentation learning algorithm driven by minimum conditional description length that we introduced in Section 5. We will try three different variants on this algorithm: one without an ITG to condition on, one conditioned on the chunked ITG, and one conditioned on the chunked ITG with categories. The first variant is completely independent from the chunked ITGs, so we will also try to create mixture models with it and the chunked ITGs.

Since the MCDL objective tends to segment large rules and count on them being recreatable when needed, many of the longer rules that would be good

Table 1: Experimental results. *Chunked* is the base model, which has categories added to produce *chunked w/categories*. *Segmented* corresponds to the second learning stage, which can be done in isolation (*only*), *mixed with* a base model, or *conditioned on* a base model.

Model	**BLEU**	**NIST**	**Categories**
Chunked ITG only	3.76	0.0119	1
Chunked ITG w/categories only	9.39	0.7481	8
Segmented ITG only	17.53	4.5409	1
Segmented ITG mixed with chunked ITG	10.23	0.2886	1
Segmented ITG mixed with chunked ITG w/categories	12.06	1.1415	8
Segmented ITG conditioned on chunked ITG	17.04	4.4920	1
Segmented ITG conditioned on chunked ITG w/categories	19.02	4.6079	8
... with iterations combined	20.20	4.8287	8
... and improved search parameters	20.93	4.8426	8

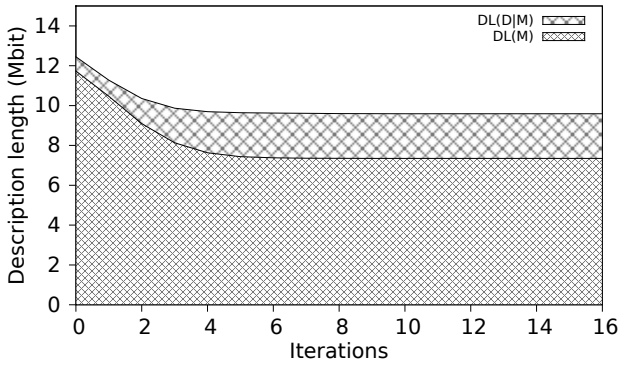

Figure 1: Description length in bits over the different iterations of segmenting search. The lower portion represents the conditional description length of the model, $DL\,(\Phi|\Psi)$, and the upper portion represents the conditional description length of the data given the model, $DL\,(D|\Phi, \Psi)$.

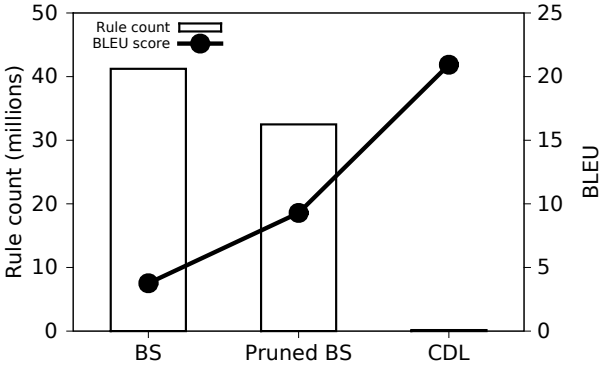

Figure 2: Rule count versus BLEU scores for the bootstrapped ITG, the pruned bootstrapped ITG and the segmented ITG conditioned on the pruned bootstrapped ITG.

to have when translating are not explicitly in the grammar. This is potentially a source of translation mistakes, and to investigate this, we create a mixture model from iterations of the segmenting learning process leading up to the learned ITG.

All the above outlined ITGs are trained using the IWSLT07 Chinese–English data set (Fordyce, 2007), which contains 46,867 sentence pairs of training data, and 489 Chinese sentences with 6 English reference translations each as test data; all the sentences are taken from the traveling domain. Since the Chinese is written without whitespace, we use a tool that tries to clump characters together into more "word like" sequences (Wu, 1999).

To test the learned ITGs, we use them as translation systems with our in-house ITG decoder. The decoder uses a CKY-style parsing algorithm (Cocke, 1969; Kasami, 1965; Younger, 1967) and cube pruning (Chiang, 2007) to integrate the language model scores. For language model, we use a trigram language model trained with the SRILM toolkit (Stolcke, 2002) on the English side of the training corpus. To evaluate the resulting translations, we use BLEU (Papineni et al., 2002) and NIST (Doddington, 2002).

7 Results

In this section we present the empirical results: bilingual categories help translation quality under the experimental conditions detailed in the previous section. The results are summarized in Table 1. As predicted the base *chunked only* ITG fares poorly, while the categories help a great deal in the *chunked w/categories*

only ITG—though the scores are not very reliable when in this low range.

The trade-off between model and data size during segmentation conditioned on the ITG with categories is illustrated in Figure 1. It starts out with most of the total description being used to describe the model, and very little to describe the data. This is the degenerate situation where every sentence pair is its own lexical rule. Then there is a sharp drop in model size with a slight increase in data size. This is where the most dramatic generalizations take place. It levels off fairly quickly, and the minor adjustments that take place on the plateau still represent valid generalizations, they just have a very small effect on the over-all description length of either the model or the data.

That the chunked ITG with split categories suffers from having too many irrelevant rules is clearly seen in Figure 2, where we plotted the number of rules contrasted to the BLEU score. Merely pruning to a threshold helps somewhat, but the sharper improvement—both in terms of model size and BLEU score—is seen with the filtering that MCDL represents.

A number of interesting lessons emerge from the results, as follows.

7.1 Minimum CDL outperforms mixture modeling

The segmenting approach works as expected (*segmented only*), essentially reproducing the results reported by Saers et al. (2013) for this style of bilingual grammar induction.

Interestingly, however, where they had success with the mixture model combining the base ITGs with the ITG learned through the segmenting approach (*segmented mixed with...*), we see a significant drop in translation quality. This may be because we have categories in our base ITG and they do not.

7.2 Category induction strongly improves minimum CDL learning

When we use the base ITGs to condition the segmenting approach, we see something interesting. The base ITG that has categories causes a sharp 1.5 BLEU point rise in translation quality (compare *segmented only* to *segmented conditioned on chunked w/categories*).

In contrast, the base ITG that does not have categories causes a slight 0.5 BLEU point *fall* in translation quality (compare *segmented only* to *segmented conditioned on chunked*).

7.3 Redundant segmental rule granularities help

As mentioned, the minimum description length objective may be theoretically nice, but it also relies on the learned ITG being able to reassemble segmented rules with fairly high fidelity at decoding time. To demand that all transduction rules are reduced to exactly a single right level of granularity may be a bit of a tall order.

Our way to test this was to uniformly mix the ITGs at different iterations though the segmenting process. By mixing the ITG after each iteration up to the one labeled *segmented conditioned on chunked w/categories*, we get the same model labeled *...with iterations combined*, which secures an additional 1.18 BLEU points.

7.4 Tuning search parameters

Lastly, for the best approach, we further experimented with adjusting the parameters somewhat. Pruning the base grammar harder (a threshold of 10^{-10} instead of 10^{-50}), and allowing for a wider beam (100 items instead of 25) during the parsing part of the segmenting learning approach, we see the BLEU score rise to 20.93.

7.5 Analysis of learned rules

A manual inspection of the content of the categories learned reveals that the main nonterminal contains mainly structural rules, segments that it could not segment further. The latter type of rules varies from full clauses such as that ' s a really beautiful dress/真是件漂亮的衣服 to reasonable translation units such as Kazuo Yamada/Ｋａｚｕｏ Ｙａｍａｄａ, which is really hard to capture because each Latin character on the Chinese side is its own individual token whereas the English side has whole names as individual tokens.

A second nonterminal category contains punctuation such as full stop and question mark, along with , sir/，先生, which can be considered as a form of punctuation in the domain of the training data.

A third nonterminal category contains personal pronouns in subject form (I, we, he, and also ambiguous pronouns that could be either subject or object form such as you and it) paired up with their respective Chinese translations. It also contains please/请, which—like pronouns in subject form—occurs frequently in the beginning of sentence pairs.

A fourth nonterminal category contains pairs such

as can/吗, do you/吗, is/吗, could you/吗 and will you/吗 — instances where Chinese typically makes a statement, possibly eliding the pronoun, and adds the question particle (吗) to the end, and where English prefixes that statement with a verb; both languages use a question mark in the particular training data we used. The main nonterminal learned that this category typically was used in inverted rules, and the other translation equivalences conform to that pattern. They include where/在哪, where the Chinese more literally translates to on/at which, what/什么 which is a good translation, and have/了, where the English auxiliary verb corresponds well to the Chinese particle signaling *perfect aspect*—that the action described in the preceeding clause is finished.

Other categories appear to still be consolidating, with a mix of nouns, verbs, adjectives, and adverbials. Chinese words and phrases typically can function as any of these, so it is possible that differentiating them may require increased emphasis on the English half of the rules.

Although the well-formed categories are few and somewhat trivial, it is very encouraging to see them emerging without any form of human supervision. Future work will expand to continue learning an even wider range of categories.

8 Conclusions

We have presented the first known experiments for incorporating bilingual category learning within completely unsupervised transduction grammar induction under end-to-end matched training and testing model conditions. The novel approach employs iterative rule segmenting driven by a minimum conditional description length learning objective, conditioned on a prior defined by a stochastic ITG containing automatically induced bilingual categories. We showed that this learning objective is superior to the previously used mixture model, when bilingual categories are involved. We also showed that the segmenting learning algorithm may be committing too greedily to segmentations since combining the ITGs with different degrees of segmentation gives better scores than any single point in the segmentation process; this points out an interesting avenue of future research. We further saw that the segmenting minimization of conditional description length can pick up some of the signal in categorization that was buried in noise in the base ITG the induction was conditioned on, leading to an ITG with much clearer categories. In total we have seen an improvement of 3.40 BLEU points due to the incorporation of unsupervised category induction.

Acknowledgements

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract no. HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the FP7 grant agreement no. 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF620811, GRF621008, and GRF612806. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, the EU, or RGC.

References

Phil Blunsom and Trevor Cohn. 2010. Inducing synchronous grammars with slice sampling. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 238–241, Los Angeles, California, June.

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. Bayesian synchronous grammar induction. In *Advances in Neural Information Processing Systems 21 (NIPS 21)*, Vancouver, Canada, December.

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A Gibbs sampler for phrasal synchronous grammar induction. In *Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, pages 782–790, Suntec, Singapore, August.

David Burkett, John Blitzer, and Dan Klein. 2010. Joint parsing and alignment with weakly synchronized grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 127–135, Los Angeles, California, June.

Colin Cherry and Dekang Lin. 2007. Inversion transduction grammar for joint phrasal translation modeling. In *Syntax and Structure in Statistical Translation (SSST)*, pages 17–24, Rochester, New York, April.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

John Cocke. 1969. *Programming languages and their compilers: Preliminary notes*. Courant Institute of Mathematical Sciences, New York University.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *The second international conference on Human Language Technology Research (HLT '02)*, pages 138–145, San Diego, California.

C. S. Fordyce. 2007. Overview of the IWSLT 2007 evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT 2007)*, pages 1–12.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pages 961–968, Sydney, Australia, July.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised ITG models. In *Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, pages 923–931, Suntec, Singapore, August.

Tadao Kasami. 1965. An efficient recognition and syntax analysis algorithm for context-free languages. Technical Report AFCRL-65-00143, Air Force Cambridge Research Laboratory.

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. An unsupervised model for joint phrase alignment and extraction. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 632–641, Portland, Oregon, June.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. 2012. Machine translation without words through substring alignment. In *50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 165–174, Jeju Island, Korea, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, Pennsylvania, July.

Jason Riesa and Daniel Marcu. 2010. Hierarchical search for word alignment. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 157–166, Uppsala, Sweden, July.

Jorma Rissanen. 1983. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, June.

Markus Saers and Dekai Wu. 2009. Improving phrase-based translation via word alignments from stochastic inversion transduction grammars. In *Third Workshop on Syntax and Structure in Statistical Translation (SSST-3)*, pages 28–36, Boulder, Colorado, June.

Markus Saers and Dekai Wu. 2011. Principled induction of phrasal bilexica. In *15th Annual Conference of the European Association for Machine Translation (EAMT-2011)*, pages 313–320, Leuven, Belgium, May.

Markus Saers, Joakim Nivre, and Dekai Wu. 2009. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *11th International Conference on Parsing Technologies (IWPT'09)*, pages 29–32, Paris, France, October.

Markus Saers, Joakim Nivre, and Dekai Wu. 2010. Word alignment with stochastic bracketing linear inversion transduction grammar. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 341–344, Los Angeles, California, June.

Markus Saers, Karteek Addanki, and Dekai Wu. 2012. From finite-state to inversion transductions: Toward unsupervised bilingual grammar induction. In *24th International Conference on Computational Linguistics (COLING 2012)*, pages 2325–2340, Mumbai, India, December.

Markus Saers, Karteek Addanki, and Dekai Wu. 2013. Combining top-down and bottom-up search for unsupervised induction of transduction grammars. In *Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-7)*, pages 48–57, Atlanta, Georgia, June.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October.

Ray J. Solomonoff. 1959. A new method for discovering the grammars of phrase structure languages. In *International Federation for Information Processing Congress (IFIP)*, pages 285–289.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)*, pages 901–904, Denver, Colorado, September.

Zhibiao Wu. 1999. LDC Chinese segmenter.

Daniel H. Younger. 1967. Recognition and parsing of context-free languages in time n^3. *Information and Control*, 10(2):189–208.

Hao Zhang, Chris Quirk, Robert C. Moore, and Daniel Gildea. 2008. Bayesian learning of non-compositional phrases with synchronous parsing. In *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 97–105, Columbus, Ohio, June.