Word Alignment of Proof Verbalizations Using Generative Statistical Models Final Report

Stephen Chong
Cornell University

May 13, 2002

1 Introduction

This project investigated the effectiveness of using generative statistical models to align semantic concepts with human-produced verbalizations of the semantic concepts. The effectiveness was measured qualitatively, and indirectly compared and contrasted with the results of (Barzilay and Lee, 2002), a previous study that uses different alignment techniques on the same data.

Aligning semantic concepts to verbalizations has applications in the construction of mapping dictionaries for a lexical chooser of a natural language generation (NLG) system, and also in the analysis of how humans convert semantic concepts into natural language.

The semantic concepts that were studied in this project were high-level mathematical proofs from the Nuprl system (Constable et al., 1986); the generative statistical models considered were the IBM models 1 and 2 (IBM-1 and IBM-2) (Brown et al., 1993), and the Hidden-Markov alignment model (HMM) described in (Vogel et al., 1996) and (Och and Ney, 2000a).

Section 2 provides some motivation and background information for this project. A survey on the literature pertaining to statistical generative alignment models is presented in Section 3, with discussion of issues pertinent to this project. Section 4 presents the methodology used in this project, focusing on the corpus used, the training and evaluation of the models, and the parameter settings for the models. Section 5 discusses the results of the evaluations performed, and Section 6 concludes.

2 Motivation

Recent years have seen an increase in the body of formal mathematical proofs. These proofs are generated by, or with the assistance of, a number of automatic theorem provers and proof assistants. These include Nuprl, Alf, Coq, OMEGA, Isabelle and HOL.

This growing body of formalized mathematics is becoming increasingly available, through online libraries and archives, such as the Nuprl Digital Library project. This is a 5 year project commenced in 2001, aiming to create an extensive digital library of many different systems' formal proofs, and to thus provide a forum for communication, and a resource for reference and education.

However, availability is not the same as accessibility—the target audience of a formal mathematics digital library may include many who have not received specific training in a given system's formalism. Indeed, one of the explicit aims of the Nuprl project is to "[empower] the lay scientist,"[1] who presumably is not a Nuprl expert.

Proof verbalization is the automatic generation of human-readable natural language proofs from formal proofs. Proof verbalization can provide a means to make these available resources accessible.

[1] From the abstract of the Nuprl Digital Library Project, available at http://www.cs.cornell.edu/Info/Projects/NuPrl/documents/Constable/bldgAbstract.html

Some recent work on the verbalization of formal proofs include (Coscoy et al., 1995), that produces "pseudo natural language" from the Coq system, the PROVERB system (Huang and Fiedler, 1997b), (Huang and Fiedler, 1997a) for verbalizing proofs from OMEGA, and (Holland-Minkley et al., 1999) and (Barzilay and Lee, 2002), both of which produce verbalizations of proofs in the Nuprl system.

Alignment is the linking of tokens or units between strings. In NLP, the term *alignment* generally means the aligning of linguistic units across two languages; *word alignment* is the aligning of linguistic units below the sentence level across two languages. Word alignment of proof verbalizations is word alignment where one language is a formal, semantic language, detailing a proof, and the other is a natural language.

The word alignment of proof verbalizations has relevance both in hand-crafted and knowledge-lean approaches to proof verbalization.

In a knowledge-lean style approach, where the verbalization problem is regarded as being a machine translation problem, the application is quite direct: the induction of a bilingual lexicon, or mapping dictionary. The mapping dictionary is used by the lexical chooser of a NLG system that verbalizes proofs.

The application of alignments to hand-crafted NLG systems lies in the need to understand what verbalizations humans produce from proofs, and how the verbalization of small semantic expressions compose together to form verbalizations of larger semantic expressions. (Holland-Minkley et al., 1999) has analyzed a corpus of proofs and their verbalizations, but primarily at the sentence level, rather than at the word level. Analyzing how verbalization works at an atomic level is useful in, for example, extending the work of (Coscoy et al., 1995) to produce more human-like verbalizations, instead of their "pseudo-natural language" verbalizations. To do this would require determining what kinds of transformations from semantics into verbalizations humans perform, and whether the implementation for displaying expressions are adequate to deal with this.

So, aligning proofs and their verbalizations at the word level has at least a couple of applications in proof verbalization. This project investigated the feasibility of using generative statistical models to produce alignments for use in proof verbalization.

3 Literature Survey

(Brown et al., 1993) introduced the 5 well-known IBM generative statistical models for word alignment. This paper is one of the first to discuss aligning words within sentences, although previous work had been done on obtaining pairs of aligned sentences from parallel corpora, using simple statistical methods.

This work is motivated by the statistical approach to machine translation, where, if we are given a sentence $\mathbf{f}$ in French (the source language) that we want to translate into English (the target language), the best translation is the English sentence $\mathbf{e}$ that maximizes $\Pr(\mathbf{e})\Pr(\mathbf{f}|\mathbf{e})$.

The IBM models are ways to estimate the probabilities $\Pr(\mathbf{f}|\mathbf{e})$. Models which estimate these probabilities are known as *translation models*, as they model how the English sentence $\mathbf{e}$ gets translated into a French sentence $\mathbf{f}$. Note that this means they model how a sentence in the *target* language get translated into a sentence in the *source* language—this inversion can easily lead to confusion.

The IBM models are generative models, in that they tell a story of how if we are given an English sentence $\mathbf{e}$, we can generate a French sentence $\mathbf{f}$. All the models use the concept of word alignment in their stories, that is, the idea that the words in the English sentence are aligned with 0 or more positions in the French sentence, and that a word in a given position in the French sentence is chosen based on which English word (if any) that position was aligned with. The alignment actually used in the IBM models is a little more restrictive: a French word can be aligned to at most 1 position in the English sentence (it may not be aligned to any position at all). The models differ in how the alignments between the two sentences are chosen. IBM-1 makes the simplest assumptions about how the alignments are chosen, and the models increase in the complexity of their assumptions up to IBM-5. A good and light-hearted introduction to the concepts involved in statistical machine translation and the IBM models can be found in (Knight, 1999).

Since its publication, a lot of attention has been focused on the IBM models—the Citeseer Research Index listed over 100 papers that reference (Brown et al., 1993). Of particular note is the John Hopkins

University's Center for Language and Speech Processing's Summer Workshop in 1999, (Al-Onaizan et al., 1999), which produced the EGYPT toolkit, a publicly available implementation of IBM-1, IBM-2 and IBM-3. The EGYPT toolkit was used in this project. Och (who was a member of the workshop) and Ney later produced GIZA++, (Och and Ney, 2000b), an extension to EGYPT that implements IBM-3 and IBM-4, and HMM (Vogel et al., 1996), (Och and Ney, 2000a). This tool was also used in this project.

HMM was introduced in (Vogel et al., 1996). Its generative story is similar to that of IBM-1 and IBM-2, but differs in the assumptions of how alignments are chosen: IBM-1 says that all possible alignments are equally likely; IBM-2 assumes[2] that when we choose where position j in the French sentence gets aligned with, this choice depends only on j; HMM assumes[3] instead that this choice depends on where we aligned position $j-1$.

When HMM is used, it generally replaces IBM-2 in training schemes. It has the same parameters as IBM-1 and IBM-2 (whereas IBM-3, IBM-4 and IBM-5 require additional parameters), which makes it easy to substitute HMM for IBM-1 or IBM-2. I assume that it *replaces* IBM-2, rather than being used *in addition* to IBM-2 because their assumptions about how alignment works conflict, and thus using the parameters found by one of the models as the starting parameters for the other would not tend to speed up the training process.

(Och and Ney, 2000a) adds some extensions to HMM, such as allowing for the spurious generation of words in the source language sentence (i.e. not every position in the source language sentence needs to be aligned with a position in the target language sentence). The implementation of this extended version of HMM is provided in the GIZA++ toolkit.

The studies (Och and Ney, 2000a), (Och and Ney, 2000b) show that using HMM as a replacement for IBM-2 resulted in better word alignment for their particular domain. Although they don't discuss what properties this model has that may explain its success (other than that it is a first-order model, while IBM-2 is a zero-order model), examining (Vogel et al., 1996) seems to indicate that HMM will outperform IBM-2 when one word in the source language is aligned with many sequential words in the target language and the alignment appears "off the diagonal." This is due to an innate assumption in IBM-2, whereby when choosing what (target language) position i to align (source language) position j to, there is a tendency towards choosing position $i = j$ in the target language sentence, i.e. a tendency towards the diagonal.[4] In the domain of proof verbalizations I do not believe there is a particular tendency for "correct" alignments be on the diagonal, and there may be a tendency for the correct alignments to have sequential runs, as the HMM model did indeed out-perform IBM-2. This result will be discussed in §5.

Are these models suitable for application in the domain of proof verbalizations? Previous applications of these models have been in aligning one natural language to another. The best results seem to be obtained when the two natural languages are structurally similar, and according to (Yamada and Knight, 2001), "it has been suspected that a language pair with very different word order such as English and Japanese would not be modeled well by [IBM-style translation models]." How successful can we expect these models to be when the source language sentences are semantic expressions rather than natural language sentences?

The models are relatively naive linguistically—they make very few assumptions about properties of the language pairs. Indeed, this is touted as being one of the key strengths of the models, that they can be applied to different language pairs provided there is a sufficient parallel corpus.[5] However, there are some assumptions that are built into the models, that may manifest themselves as difficulties in some natural language pairs, but prove to be more problematic in our domain of proof verbalization.

For example, all IBM models and HMM are limited in that alignments can only be many-to-one, from the source language to the target language (i.e. a word in the source language French, can be aligned with at most one word in the target language English, while a word in the target language English can be aligned with possibly many French words). This short-coming has shown itself to be problematic in some language pairs if the sentence lengths are vastly different, or if the corpus contains many idioms, which require many-to-many alignments. For example, German frequently uses compound words, where English does not. This results in German sentences having fewer words than their translations. This is not a problem if translating

[2] Actually, it is assumed that it also depends on the lengths of the target and source sentences, for normalization reasons.
[3] Again for normalization purposes, it actually also depends on the length of the target (English) sentence.
[4] See §3.1, (Vogel et al., 1996).
[5] See §1.7 in (Brown et al., 1993).

from English into German, but is a problem if translating in the other direction.[6] In our project, the source language is the Nuprl proofs. Although the Nuprl proofs tend to have more tokens than the English verbalizations, the alignments constructed by the Nuprl experts as part of the evaluation tend to be very sparse in their alignments. We may thus expect to see better results by swapping the source and target languages. See §5 for further discussion.

Apparently there are some techniques to help alleviate the many-to-many alignment problem, by introducing phrases (Och et al., 1999), (Wang, 1998). Some of these techniques seem to be direct solutions, in terms of having a generative model that permits many-to-many alignments, and are generalizations of the IBM models. Other techniques seem to consist of ways of avoiding the problem, such as identifying phrases, and adding them directly to the vocabulary, i.e. regarding the phrase as a single word.

(Yamada and Knight, 2001) presents a statistical translation model that operates not on sentences in the target and source languages, but on a target language parse tree and a source language sentence. That is, the statistical machine translation is modified, so that instead of the probabilities $\Pr(\mathbf{f}|\mathbf{e})$ needing to be estimated, where $\mathbf{f}$ is a sentence in the source language (French) and $\mathbf{e}$ is a sentence in the target language (English), the probabilities $\Pr(\mathbf{f}|\mathcal{E})$ are estimated, where $\mathbf{f}$ is a sentence in the source language and $\mathcal{E}$ is a parse tree in the target language.

The model they use is a generative one that allows alignments to be made between words in the source and target languages. The "story" that is told by this model is that the target language tree can have certain operations performed on it, including the re-ordering of children, the insertion of new nodes, and the translation of leaf nodes (target language words) into source language words. This model is interesting to us, in that if we regard the Nuprl proofs as the target language, we notice that Nuprl proofs have a natural, unambiguous tree structure—the structure of the Nuprl term. Although they call their model "syntax-based," I don't believe there is anything to prevent it being used with general tree structures, or for that matter, regarding the Nuprl structural tree as being a syntax tree. Thus, we could perhaps apply their model to our domain, and perhaps take advantage of the structural information that is readily available with the Nuprl proofs. By representing the structure of the Nuprl proof in a tree, we could remove the need for parentheses to appear as tokens, or parts of tokens. This could help alleviate some of the issues involved with the tokenization of mathematics, as discussed in §4.1.

The results of their experiments with this model show that it compares well against the IBM-5, with Japanese and English as the source and target languages respectively. This bodes well for hopes that this model will also perform better than the IBM models in our domain.

(Barzilay and Lee, 2002) presents a system for producing verbalization of Nuprl proofs that makes use of a mapping dictionary from semantic concepts to text templates. The mapping dictionary is induced from the same corpus as this project will use, through innovative application of multi-sequence alignment techniques developed for use in computational biology. The techniques they use do not directly produce alignments from Nuprl proof steps to verbalizations; instead, the multi-sequence alignment techniques take several different verbalizations of the same Nuprl proof step, and align these verbalizations with each other. In a process that applies multi-sequence alignment techniques several times, templates are ultimately produced that map from Nuprl predicates to natural language templates, with slots for inserting verbalizations of terms.

In theory, the induction of the mapping dictionary could have been achieved through the application of standard machine translation methods, such as the statistical alignment models being considered in this project. However, Barzilay and Lee point out some potential problems with applying these methods. First, they point out that there is frequently a large discrepancy in what information different authors actually present in the text they produce from a proof: "given the same semantic input, different authors may (and do) delete or insert information in the text they produce." The concern is that a single verbalization will not be sufficient for the methods to extract the real correspondences between the verbalization and the semantic information. However, this objection would seem to also apply to many corpora to which statistical machine translation methods have been applied. For example, I found the following example from the Hansard bilingual corpus:

English: My wife, Diana, and I were happy to welcome Her Majesty the Queen...

[6] See (Och and Ney, 2000a).

French: Ma femme et moi avons eu la joie d'accueillir Sa Majesté la Reine...

The English version of the sentence contains the name of the speaker's wife; the French does not. I believe many bilingual corpora have non-literal translations, which create substantial difficulty for statistical machine translation, but show that the insertion or omission of information in translations is not unique to the domain of proof verbalizations, and appears to be present in at least some domains where statistical machine translation methods have been applied with relative success. Nevertheless, examination of the alignments from Nuprl proof steps to English verbalizations constructed by Nuprl experts show that there seems to be a phenomenal amount of information insertion and omission occurring in this corpus. See §5 for more details.

The second potential problem that is pointed out, is that "a single verbalization certainly fails to convey the variety of potential linguistic realizations of the concept that an expressive lexical chooser should have access to." While this is indeed true, wouldn't this only be a problem if the corpus contains only a single verbalization of a concept? Statistical machine translation relies on corpora large enough to have multiple occurrences of tokens in the source and target languages. Admittedly, in the domain of Nuprl proof verbalizations, the inference of tactic translations will be complicated by the different verbalizations of different arguments to tactics, which will further exacerbate the sparse data problem.

4 Methodology

This section details the methodology that was used to prepare the corpus of Nuprl proof verbalizations, apply the generative statistical models IBM-1, IBM-2 and HMM to the corpus and to evaluate the results. It discusses issues that were encountered during these tasks, and discusses the impact these issues may have had on the results.

4.1 Corpus

The corpus of Nuprl high-level proofs and their verbalizations that was used derived from data originally gathered as part of the study by Holland-Minkley et al. (1999). This data was used by Barzilay and Lee (2002) and formed the core of their training data, which consisted of 30 different Nuprl proofs, and 83 human written verbalizations of these proofs. The verbalization texts were divided into pieces corresponding to individual proof steps as pre-processing performed for their study. One proof step roughly corresponds to one sentence in a natural language verbalization, as shown in (Holland-Minkley et al., 1999). A copy of this pre-processed corpus of single proof steps and their corresponding verbalizations was obtained for use in this project.

Preprocessing: Some amount of pre-processing was required on the corpus as received from Barzilay and Lee.

First, their corpus consisted of pairs of Nuprl proof steps and sets of verbalizations of the given proof step; the generative statistical models used in this project required sentence pairs, and as such a pair $< \mathbf{e}, \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\} >$ from Barzilay and Lee's corpus was converted into the pairs $< \mathbf{e}, \mathbf{f}_1 >, < \mathbf{e}, \mathbf{f}_2 >, \ldots, < \mathbf{e}, \mathbf{f}_n >$, where $\mathbf{e}$ is a Nuprl proof step, and $\mathbf{f}_i$ is a proof verbalization of $\mathbf{e}$. After this conversion, the corpus consisted of 338 sentence pairs, from 134 distinct Nuprl proof steps.

Second, both the Nuprl proof steps and the English verbalizations used a non-standard 8-bit font, to display certain mathematical notation, such as $\mathbb{Z}$ and $\mathbb{N}$. This raised several issues. The corpus was obtained via an HTTP transfer of text files which used the non-standard font. During the HTTP transfer, some spurious control characters appear to have been introduced into the corpus. These spurious characters had to be removed. Also, displaying the non-standard font outside of the Nuprl environment was problematic—a substantial amount of work was required to enable the Nuprl proofs to be displayed correctly in the Cairo alignment visualization tool of the EGYPT toolkit.[7] Indeed, the extended font set seems to have had some impact on the creation of the corpus itself: the English verbalizations of proofs have a wide variety of notation, including copy-and-pasted Nuprl text, direct English translations of symbols, and ASCII versions of the symbols. For example, the bidirectional arrow symbol $\Leftrightarrow$ was sometimes copied and pasted from

[7] Cairo was used by the Nuprl experts during the formal evaluation

the Nuprl directly into the English verbalizations, sometimes translated as `if and only if` or `iff`, and sometimes the characters `<==>` were used. Other examples include the use of `Z` for $\mathbb{Z}$ and the use of `div` for $\div$. Other examples abound in the corpus.

Third, a number of spelling mistakes existed in the English verbalizations of the proofs, for example `immediatelly`, `arithemetic` and `hypothetheses`. Due to the small size of the corpus, it was practical to examine all distinct tokens, and determine if any of them were typographic errors, and correct them. This was done, resulting in the correction of 7 tokens. In previous work, such as (Brown et al., 1993), rare tokens have been replaced in the corpus with a special *unknown* token, in an attempt to remove typographic errors. The ability to manually inspect all tokens obviates such a need. However, replacing rare tokens with a special token was tried during the experiments, to see if the models performed better with smaller vocabularies.

Tokenization: There were two issues involved with the tokenization of the corpus.

First, the tokenization of the English verbalizations was inconsistent: some of the English parts of the corpus had certain punctuation split out as separate tokens, such as colons, periods and commas. However, other parts of the corpus did not. For example, there were 6 occurrences of `integer` immediately followed by a comma, and 3 occurrences of `integer` followed by 1 or more whitespace characters, and then a comma. In total, there were 117 occurrences of a comma, colon or period that was not split out as a separate token. A number of these occurrences were part of mathematics quoted in English, and were tokenized according to the guidelines below. For the remainder, whitespace was inserted to ensure that the comma, colon or period was treated as an individual token.

The second issue was how to tokenize the Nuprl proof steps. The proof steps are in a formal language, rather than a natural language, and include far more punctuation and structure than natural language does—there are a plethora of parentheses, colons, and arcane mathematical symbols[8].

There is a simple way to tokenize the Nuprl proof steps: just use the existing whitespace to split the tokens. However, we may be able to do better than this, for the following reason. A criticism of the IBM models is that they work best for similar language pairs. Indeed, they do seem to perform better if the languages are similar. So it is in our interest to attempt to tokenize the Nuprl proof steps so that the tokens are of approximately the same semantic granularity as the English tokens, if possible.

Now, the English verbalizations frequently quote some of the mathematics from Nuprl, either directly or paraphrased. The key point is that the English verbalization contains mathematical notation, not just natural language. If we could ensure that the mathematics quoted in English has the same tokenization as the mathematics in the Nuprl proof steps, then we are helping to ensure that the tokens in the source and target language have approximately the same semantic granularity.

This task is complicated by the fact that mathematics in English is not necessarily quoted with the same whitespace as in the Nuprl. For example, the corpus contains a proof step with the substring `x = y`, and a corresponding verbalization containing the substring `x=y`. Of course, the English verbalizations are not consistent amongst themselves, and there is a verbalization containing the substring `x = y`.

An interesting point to note is that the English verbalizations frequently quote mathematics from the Nuprl proof step, but generally only quote small chunks of mathematics. In particular, with very few exceptions, the mathematics quoted in the English verbalizations do not seem to have deep levels of parenthesis nesting. About 45 of the 338 verbalizations have nested parentheses, of which 14 have a nesting depth of 3. None have a depth greater than 3. Contrast this with all of the Nuprl proof steps having nested parentheses of at least depth 3, and 205 of them having depth 4 or more. This is some evidence in support of the possible success of Yamada and Knight's syntax tree translation (see §3 above).

Based on this fact, as an initial attempt to tokenize the Nuprl mathematics into approximately the tokens used in the English verbalizations, I separated out the parentheses, ensuring that if a token contained parentheses, then it contained nothing but parentheses. I also ensured that if two tokens consisting of nothing but parentheses were adjacent, then they were merged into a single token. However, this resulting in an immense increase in the number of tokens in the Nuprl proof steps. There were 3484 occurrences of `()`, 2670 of `)`, 951 of `))` and 598 of `(((`. There were fewer occurrences of some more exotic parenthesis combinations, such as `())`. By comparison, the most frequent non-parenthetical tokens were the colon (728 occurrences) and $\mathbb{Z}$ (594 occurrences). This increase in the number of tokens exacerbated the sentence length problem (see below), and due to the fact that most of the tokens were parentheses, resulted in very poor word alignments.

[8] $\mathbb{Z}^{-0}$ baffled me for quite a while.

I next attempted to remove all parentheses from the Nuprl proof steps completely. While this did in fact produce some reasonable word alignments, I decided against pursuing this tokenization, as it means the tokenized Nuprl proof steps are no longer an accurate reflection of the semantics. For example the expression `x · (y + c)` with the parentheses removed has a very different meaning: `x · y + c`.

In the end, a decision was made to leave the tokenization of the mathematics in the English verbalizations alone, and to apply some heuristics to the tokenization of the Nuprl mathematics only, to attempt to make them tokenized more like the English tokenization. The reason for this decision was two-fold: The Nuprl proof steps have much more consistent structure, and are thus much easier to re-tokenize automatically with regular expressions. Also, the use of mathematics in the English verbalizations tended to be more compact that the Nuprl mathematics, i.e. less whitespace, and thus fewer tokens. Reducing the number of tokens in the Nuprl proof steps helps to alleviate the sentence length problem (see below).

Sentence length: Without having inspected the corpus, I was expecting that the English verbalizations would be longer (in terms of number of tokens, for some "reasonable" tokenization) than the corresponding Nuprl proof step, as I assumed that mathematical notation would be concise and succinct, while the English verbalizations would be verbose and contain a lot of linguistic "glue," to make the sentences comprehensible, and the proof verbalizations as a whole cohesive. This expectation turned out to be incorrect: in general the Nuprl proof steps had more tokens than the English verbalization. This is partly a product of the representation of Nuprl proof steps, which contains a lot of repeated information: each proof step has a list of its assumptions, and may produce several sub-goals, each of which will have most of the same assumptions. Thus, while most of the proof steps in a Nuprl proof may contain the assumption `a:`$\mathbb{Z}$, the English verbalization will only state once that `a` is an integer.

4.2 Parameters

A number of different parameters were considered for the experiment.

- Tokenization

  As discussed above, how the corpus is tokenized is not a trivial issue. A number of different tokenizations were examined in an ad hoc manner, and one selected for more detailed examination.

- Remove rare words.

  Rare words can be removed from the corpus, during pre-processing of the corpus, and replaced with a special *unknown* token. There are only two settings for this parameter: removing the rare words, or not removing them.

- Training schemes.

  In previous work using IBM-1 to IBM-5 and HMM, parameter values found by simpler models are used to initialize the parameters for more complex models ((Brown et al., 1993), (Och and Ney, 2000b)). For example, IBM-1 can be trained for 5 iterations, and the parameter values on the final iteration used to initialize the parameters for the Hidden-Markov model, which is trained for 5 iterations, and its final parameter values used to start the training of IBM-4, which is run for 8 iterations. This training scheme would be written $1 \rightarrow \text{HMM} \rightarrow 4$.

  The following training schemes were used in this project:

  - 1
  - $1 \rightarrow 2$
  - $1 \rightarrow \text{HMM}$

  Each model in a training schedule was trained for 50 iterations.

- Switching Target and Source Languages.

  Given a parallel corpus, it is possible to switch which language is the target language, and which is the source. This is mainly of interest due to the restrictions on the alignments that can be produced

by the IBM and HMM models. In particular, the alignments produced by these models cannot have many-to-many relationships, only many-to-one from the source (e.g. French) language to the target (e.g. English) language.

So, for a given tokenization of the corpus, there are 12 distinct parameter settings.

4.3 Training and Evaluation

For each different parameter setting, the specified models were trained for 50 iterations. Due to the small size of the corpus, despite the high number of iterations, total training time was not large. This number of iterations is high. Literature which makes use of these models, such as (Yamada and Knight, 2001), use much lower number of iterations. This combined with the small corpus leads me to think that I may have risked over-specializing on corpus. However, I have not lost sleep over this. While I know that some unsupervised learning methods, such as clustering where training decides the number of clusters, can be over-trained, I am not entirely sure if it is possible to over-train a model which uses unsupervised learning like this.

Ideally a gold standard would have been used for evaluation. A gold standard is a set of reference data, which is some subset of the corpus that has been manually aligned. A gold standard as a means of evaluation has several advantages: objective measurements can be obtained, and experts' time is not required with each new data set. Evaluation with a gold standard is the recommended approach in (Ahrenberg et al., 2000). (Melamed, 1997) is one example of such a gold standard.

This project does not use a gold standard, primarily due to the enormity of the task of constructing a gold standard. As Melamed's work shows, it is time-consuming and difficult.

For the chosen tokenization, there were 12 different experiments, corresponding to the 12 possible combinations of parameter settings. Due to limited availability of Nuprl experts, it was not possible to perform a full appraisal on each of these 12 experiments, let alone all the experiments that would result if several different tokenizations were considered. Instead, I performed an informal evaluation on each of the 12 experiments, and what I regarded as the best 3 were then submitted to a fuller evaluation by 3 Nuprl experts.

My informal evaluations involved a manual inspection of the Viterbi alignments produced by the models for a small subset of the corpus. I informally judged the quality of the alignments in a similar manner to the way the evaluators judged the quality of alignments during the formal evaluation.

The formal evaluation of the experiments made use of 6 randomly selected proof steps from the corpus. There were two parts to the formal evaluation of the experiments.

In the first part, the Nuprl experts were asked to construct alignments from Nuprl proof steps to the corresponding verbalizations[9]. More precisely, they were asked to connect a Nuprl token to an English token if they felt that the Nuprl token in some way contributed to the existence of the English token. The evaluators were allowed to construct many-to-many alignments if they choose to. The evaluators performed this task on three of the six test proof steps. All evaluators used the same three Nuprl proof steps.

The purpose of this stage was two-fold: it gave me some "correct" alignments for some proof steps and verbalizations, which provides me with a basis to compare how well the generative statistical models can do, and it also ensured the evaluators had given some thought as to what constitutes a correct alignment before they started the second part of the evaluation.

In the second part of the evaluation, the Nuprl experts were presented with an alignment of a Nuprl proof step and a corresponding verbalization, and asked to judge each alignment of a Nuprl token as *Okay*, *Not sure* or *Wrong*. This follows the methodology of (Yamada and Knight, 2001). The alignments with which they were presented were the Viterbi alignments produced by the models trained with the 3 chosen parameter settings. There were thus 18 such alignments that the experts were required to evaluate.

The experts were able to examine the Viterbi alignments through the Cairo alignment visualization tool, provided with the EGYPT toolkit. This tool allowed them to easily highlight a token, in either English or Nuprl, and see which tokens in the other language it was aligned with.

[9]This evaluation was suggested by Eric Breck during the CS674 project presentations.

I was originally going to perform a third evaluation, which was intended to give some indication of the potential of using the generative statistical models for mapping dictionary induction. This evaluation was to be in lieu of a direct comparison with (Barzilay and Lee, 2002), as it was determined that the outputs of their system were not close enough to the alignments produced by the generative statistical models to have any meaningful direct comparisons.

The ethereal third evaluation would have constructed a number of tactic translations, by examining Viterbi translations, and removing any tokens in the English verbalizations that were aligned with the tactic's arguments. The remaining English tokens would then constitute the translation of tactic, with the removed tokens being the slots to insert the verbalizations of the tactic arguments. I thus retrieved the 14 Nuprl proof steps and verbalizations in the corpus in which the `Rewrite` tactic occurred (6 distinct Nuprl proof steps), took the Viterbi alignments produced by each of the 3 models on each of the 14 Nuprl proof steps and alignments, and constructed a template from each of these Viterbi alignments. However, upon examining the templates thus produced, I made an executive decision that they were of too poor a quality to subject the evaluators to. Some examples of these abortive tactic translations can be found in the appendix.

Possibly a method that made more sophisticated use of using the Viterbi alignments could produce a tactic translation of higher quality. However, such a method is not immediately apparent, and due to time constraints, I was unable to develop and evaluate such a method.

5 Results

During the informal evaluations, I examined the Viterbi alignments produced by the models trained by the 12 distinct parameter settings for the chosen tokenization. In judging Viterbi alignments, I considered one Viterbi alignment better than another if it contained more correct token alignments and/or fewer incorrect token alignments. I informally did what the Nuprl experts did in the second part of the formal evaluation.

During the course of this examination, I decided that the 3 parameter settings that had Nuprl as the source language, and did not replace rare words in the corpus with a special *unknown* token produced the best Viterbi alignments. That is, the experiments chosen for further investigation were:

- Nuprl source, keep rare words, training scheme: 1
- Nuprl source, keep rare words, training scheme: $1 \rightarrow 2$
- Nuprl source, keep rare words, training scheme: $1 \rightarrow$ HMM

One of the key characteristics of the alignments produced with English as the source language, is that many of the Nuprl tokens are not aligned. The reason for this is that the IBM and HMM models allow many-to-one relationships only from the source language to the target language. Thus, if English is regarded as the source language, then an English token can be aligned with at most one Nuprl token. As the Nuprl sentences are in general longer than the English sentences, this means that many Nuprl tokens are left unaligned. During the informal evaluations, I regarded this as a deficiency, for two reasons. First, having only a few alignments highlighted that the fact that most of the alignments that did exist were not correct. Second, I was expecting the alignments to look somewhat similar to alignments between two natural languages, where an unaligned token is the exception rather than the rule.

However, in light of the alignments constructed by the Nuprl experts, in which most of the tokens, both English and Nuprl, are left unaligned, this perceived deficiency may in fact be a virtue. Having most of the Nuprl tokens unaligned would result in a better agreement with the alignments constructed by the Nuprl experts, and thus presumably a better evaluation by the Nuprl experts.

While having English as a source language may result in better evaluation of the alignments, it is not however necessarily the best option for the ultimate goal of constructing mapping dictionaries. Indeed, the naive inference of tactic translations discussed in §4.3 would fail dismally in this case, as many of the templates inferred would be the complete verbalization, and have no slots for the insertion of arguments! Also, it is not clear that if English was regarded as the source language when templates are inferred, that we would always be able to successfully verbalize a Nuprl proof, even if that Nuprl proof used nothing but

	Alignments agreed on by		
	1 Expert	2 Experts	3 Experts
Verbalization 1	11	10	6
Verbalization 2	6	2	3
Verbalization 3	6	8	21

Table 1: Agreement between the alignments constructed by the Nuprl experts.

	Exp. 1	Exp. 2	Exp. 3	Vit. 1	Vit. 1 $\rightarrow$ 2	Vit. 1 $\rightarrow$ HMM
Verbalization 1 - Nuprl (33)	0.15	0.18	0.18	1.00	1.00	0.91
Verbalization 1 - English (21)	0.14	0.48	0.57	0.52	0.62	0.29
Verbalization 2 - Nuprl (29)	0.03	0.07	0.17	1.00	1.00	0.93
Verbalization 2 - English (8)	0.13	0.38	0.63	0.88	1.00	0.50
Verbalization 3 - Nuprl (19)	0.42	0.79	0.53	0.95	0.95	0.95
Verbalization 3 - English (17)	0.18	0.29	0.29	0.41	0.53	0.29

Table 2: Proportion of tokens that were aligned. Numbers in parenthesis indicate the number of tokens.

tactics and terms that occurred in our training corpus. The semantics that the tactic templates thus inferred would be able to cover may not be the correct Nuprl semantics.

We would have similar reasons to be concerned about the templates induced when rare tokens are removed from the corpus. Removing the rare tokens means that the tactic translations ultimately inferred may not cover the actual Nuprl semantics. In addition, the alignments that were produced by models trained on such a corpus were not promising. Due to relatively high co-occurrence of the *unknown* token in the Nuprl and the English verbalization, in many of the experiments the *unknown* token is frequently aligned to other *unknown* tokens, and quite often these *unknown* tokens have very different semantic meaning in the context. In other experiments, the most likely token that the *unknown* token will be aligned with is one of a number of very frequent tokens, such as `ℤ)`, which in general is not correct. I'm not sure which is worse.

For these reasons, the parameter settings that removed rare words, and that used English as the source language were not used in the formal evaluation.

The formal evaluation was in two parts, as described in §4.3. The first part of this formal evaluation involved the Nuprl experts constructing correct alignments for 3 proof steps and corresponding verbalizations. The second part was an evaluation of the 3 models chosen by the informal evaluation.

Table 1 summarizes the agreement between the alignments constructed by the Nuprl experts. Each token-to-token alignment constructed by each of the experts was agreed on by 1, 2 or 3 of the experts. For example, 2 of the Nuprl experts connected the Nuprl token `(RWH` with the English token `replace` in Verbalization 1. This increased the number of alignments agreed on by 2 experts in Verbalization 1 by 2.

The alignments constructed by the Nuprl experts have a reasonable degree of agreement. Most of the token-to-token alignments the experts constructed had two or more of the experts agreeing. This gives us some degree of confidence that these manually constructed alignments are in some sense correct.

Table 2 shows the proportion of tokens in each of the three verbalizations that received an alignment to another token, in the alignments constructed by the experts, and the Viterbi alignments of the same sentences produced by the three models being evaluated. Thus, the first expert aligned 14% of the 21 English tokens in Verbalization 1 with some Nuprl token.

All the alignments constructed by the experts have the property of being sparse—rarely in their alignments did more than half of the Nuprl tokens get aligned with an English token; and rarely did more than 60% of the English tokens get aligned with a Nuprl token. There are a couple of possible explanations why this should be the case.

One possibility is that in the domain of proof verbalization, there does tend to be a large amount of information insertion and omission. The large proportion of English tokens that are not aligned with Nuprl

		Training Scheme		
		1	$1 \rightarrow 2$	$1 \rightarrow$ HMM
Sentence:	1	0.12	0.12	0.13
	2	0.15	0.17	0.22
	3	0.18	0.13	0.19
	4	0.15	0.15	0.20
	5	0.11	0.07	0.16
	6	0.18	0.16	0.39
Avg over all 6 sent.		0.15	0.13	0.21

Table 3: Average scores of the expert-evaluated Viterbi alignments.

tokens signals a lot of insertion; the large proportion of Nuprl tokens that are not aligned with English tokens signal a lot of omission. This possibility is consistent with the fears of Barzilay and Lee regarding generative statistical models, and also consistent with the observation made earlier that the Nuprl proof steps tend to repeat a lot of information that is verbalized in English only once.

A second possibility is that the corpus contains a sizeable number of pairs of proof steps and verbalizations, where the verbalizations do not actually correspond to the proof step. The task of splitting up the complete verbalization into pieces corresponding to individual proof steps is itself an alignment task, and thus not necessarily straightforward. This splitting of the verbalizations was done by Barzilay and Lee as part of their study, using "dynamic programming based on the number of symbols common to both the step and the verbalization".[10] The success of their study belies this possibility, although perhaps their techniques are more robust to mismatches between proof steps and verbalizations than the generative statistical models, as they only align proof steps to verbalizations quite late in their process.

The alignments produced by the generative statistical models investigated here are not nearly as sparse. The second half of Table 2 clearly shows this. This is another example of assumptions that are implicit in the models. This tendency of the models to not produce NULL alignments is problematic in this domain, given the sparsity of the "correct" alignments constructed by the Nuprl experts.

In the second part of the formal evaluations, the 3 Nuprl experts examined Viterbi alignments, and graded each alignment as *Okay* (1.0 point), *Not sure* (0.5 points) or *Wrong* (0 points). The appendix contains some examples of Viterbi alignments from experiments with a variety of parameter settings. The experts were provided with the Cairo alignment visualization tool, which allowed them to select a token and see clearly which tokens it was aligned with.

Table 3 shows the average scores of the alignments. Each expert's grading of the token-to-token alignments for a sentence was summed, and divided by the total number of tokens in the Nuprl sentence, thus producing a number between 0 and 1. The average of the 3 expert's scores appears in the table.

None of the 3 models evaluated did particularly well. The low scores are primarily due to the sparsity of the correct alignments, versus the tendency of the models to not produce NULL alignments, as discussed above. However, of the three models evaluated, the training scheme $1 \rightarrow$ HMM did the best. As discussed in §3, the HMM model was designed with a tendency to align sequential words in the target language. This can clearly be seen in the example alignments in the appendix. The alignments constructed by the Nuprl experts often show this behavior, for example aligning the Nuprl token `BHyp` with the sequence of English tokens `use the assertion`. This corresponds to the expansion of a Nuprl semantic unit into an English phrase. I was expecting this expansion to hold true for most of the Nuprl to English alignments, and was thus initially expecting the English verbalizations to be much longer in length than the Nuprl proof steps. The fact that the English verbalizations are not generally longer than the Nuprl proof steps may limit the success of the HMM model in this domain, even though it produces the most "correct" alignments of the models evaluated.

[10] See §4.1, (Barzilay and Lee, 2002)

5.1 Future directions

There are a large number of possibilities for further investigating the potential of using generative statistical models in the domain of proof verbalization.

A number of extensions could be made to this project. In particular, a more thorough evaluation of the data already obtained would provide more insights. For example, the models that were trained with English as the source language were not formally evaluated, and may prove to be more successful than the formally evaluated models. As mentioned above, I believe they are closer to the alignments constructed by the Nuprl experts, although they may not prove to be any better when used for the construction of tactic translations.

Another extension, which would be very simple to implement and would hopefully improve the quality of the alignments, is to make explicit the fact that when Nuprl is quoted in the English verbalization, these quoted tokens should align with the Nuprl tokens. It is very easy for humans to see that if the English verbalization contains some mathematics token, such as `¬b≤0`, and the same token occurs in the Nuprl, then they most likely should be aligned. The system cannot do this, as it does not recognize the identity of the source language and target language tokens. The GIZA++ toolkit allows for the addition of an explicit dictionary, that gives translations of tokens in the source language to tokens in the target language. We could add all unique Nuprl tokens to this dictionary, and specify that the translate to the identical token in English. This, possibly combined with further work on the tokenization of mathematics in Nuprl and English, would allow the models to make some of the alignments that humans perceive as evident.

Also useful would be clarifying if it is possible to over-train the IBM and HMM models, and if so, if I used an excessive number of iterations in my training, given the size of my corpus.

This project is based on the hope that aligning Nuprl proofs and English verbalizations can ultimately be used to produce tactic translations that will correctly cover the Nuprl semantics. Questions that haven't as yet been raised are if the Nuprl proofs are actually correct representations of the semantics, and if the Nuprl proofs present the proofs in a form that is amenable to alignment with the end goal of tactic translation. The Nuprl proofs in the corpus are not the semantics as such, but rather the "sugaring" of the semantics. For example, while the real semantics in Nuprl may be closer to $\forall x : \mathbb{N}.\ \forall y : \mathbb{N}.\ \forall z : \mathbb{N}.\ x + (y + z) \geq 0$, it could be sugared as `∀ x,y,x:ℕ. x+y+z ≥ 0`. The sugaring process may introduce ambiguity and abbreviations. Possibly using different sugarings of the Nuprl semantics in the corpus may produce better alignment results, and better tactic translations.

Given that Barzilay and Lee's knowledge lean approach did produce successful tactic translations, it is interesting to compare these two approaches, and determine if there are any fundamental assumptions made by the generative statistical models that may preclude them from performing well in this domain. I believe that the paradigm of aligning the source language to the target language with the restriction of allowing many-to-one alignments in only one direction may not be appropriate for this domain. The alignments constructed by the Nuprl experts are some evidence towards this. Further investigation of what constitutes a correct alignment in this domain may help clarify the issue.

6 Conclusion

The results obtained from both the formal and informal evaluations are not promising for the potential of using generative statistical models to infer tactic translations. The alignments produced by these models are not very successful when compared with alignments constructed by Nuprl experts, or when evaluated by Nuprl experts. The ultimate goal of producing tactic translations using these techniques does not look so rosy either.

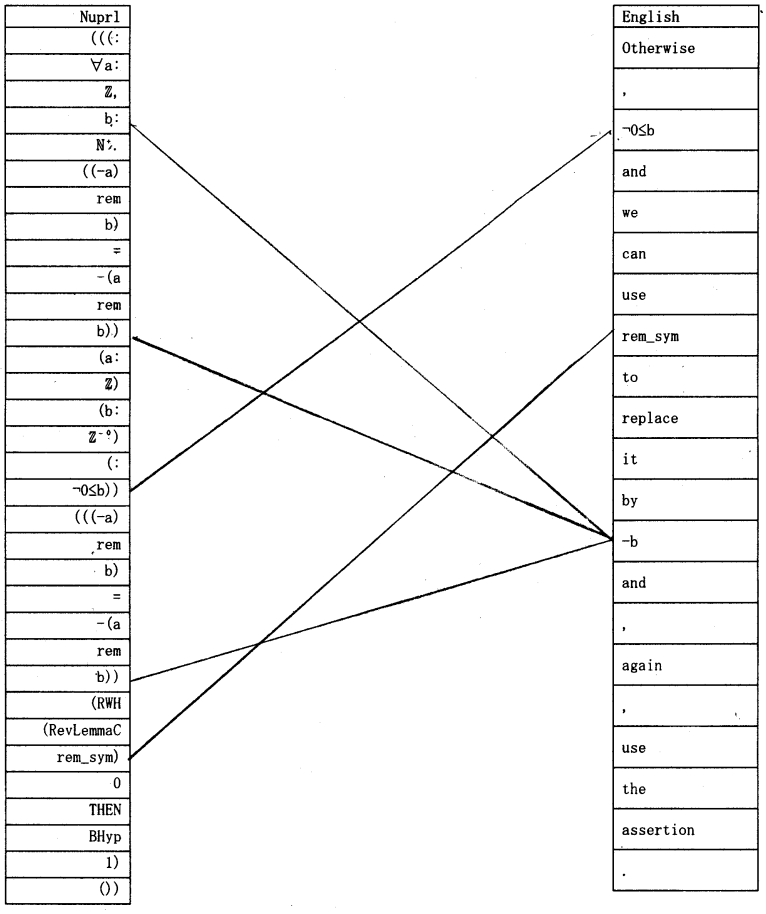
However, there are a number of avenues that could still be explored.

Proof verbalization is an important task and despite my efforts in this project, remains a difficult one.

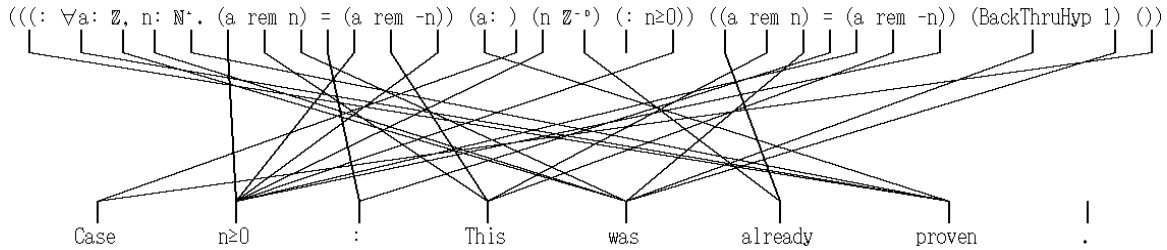
7 Appendix

7.1 Example "Correct" Alignment

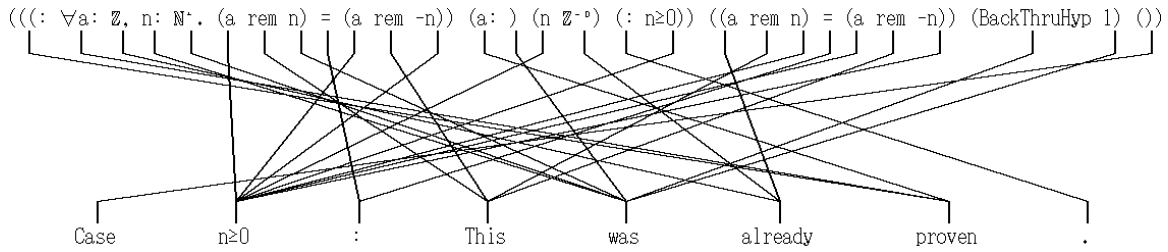
The following is an example of a "correct" alignment constructed by a Nuprl expert during the first part of the formal evaluation.



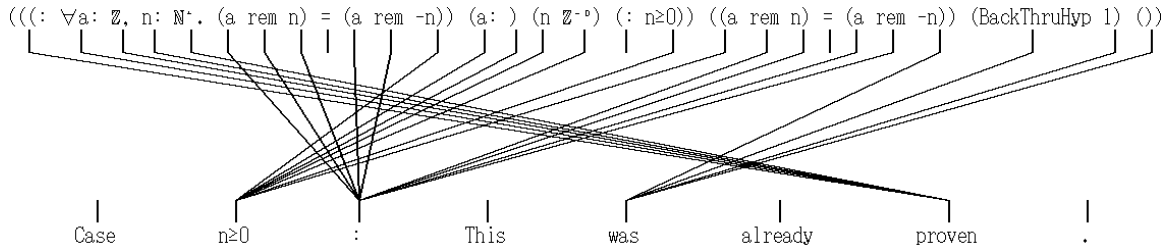
7.2 Example alignments



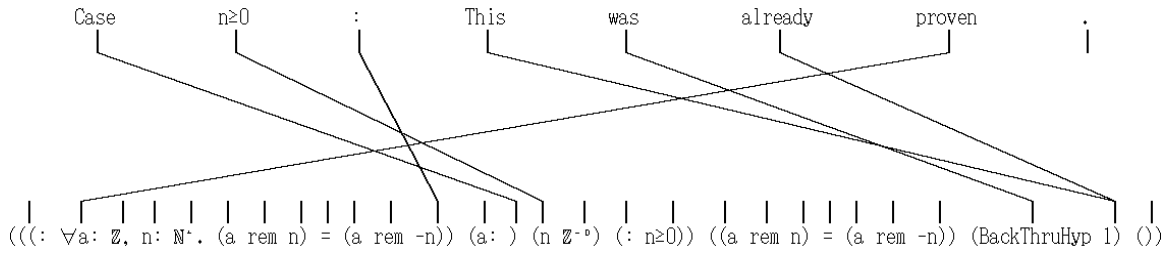
Keep rare tokens, Nuprl as source, Training scheme: 1



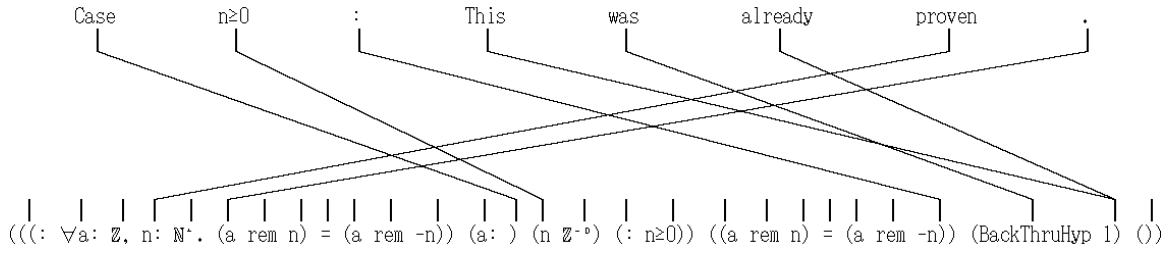
Keep rare tokens, Nuprl as source, Training scheme: $1 \rightarrow 2$



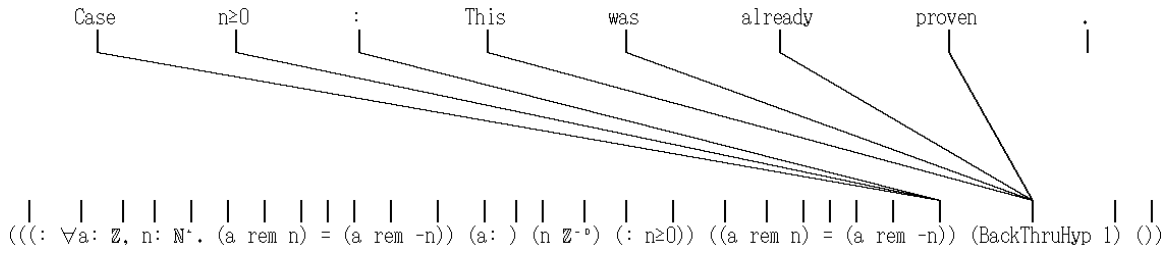
Keep rare tokens, Nuprl as source, Training scheme: $1 \rightarrow$ HMM



Keep rare tokens, English as source, Training scheme: 1



Keep rare tokens, English as source, Training scheme: $1 \rightarrow 2$



Keep rare tokens, English as source, Training scheme: $1 \rightarrow$ HMM

7.3 Unevaluated Tactic Translations

The following are some of the tactic translations induced by the Viterbi alignments of proof steps containing the `Rewrite` tactic. These tactic translations were considered too poor quality to subject to a formal evaluation. The `???` indicate tokens that were aligned with the Nuprl tactic arguments. The models

that were used to produce these tactic translations were the same three that were evaluated in the formal evaluation. The method used to produce these templates from the Viterbi alignments of the models is described in §4.3.

- Therefore, by div_rem_sum, a=(adivn)*n+(aremn) so we ??? conclude (AREMN) + (A÷N) ??? (A÷N) ·N.
- Therefore, by div_rem_sum, ??? so we can conclude (AREMN) + (A÷N) ??? (A÷N) ·N.
- According ??? the theorem div_rem_sum ???
- ??? to the theorem div_rem_sum ???
- Similarly ??? (A÷N) ·N
- ??? ??? (A÷N) ·N
- ??? A<N+ (A÷N) ·N
- By rewriting it using the ??? and add_mono_wrt_lt lemmas and ??? the result arithmetically , we can turn it into WELLFNDI (...N;X,Y.X>Y)
- By rewriting it using the ??? and ??? lemmas and ??? the result ??? , we can ??? it into ??? (...N;X,Y.X>Y)
- By rewriting it using the minus_mono_wrt_lt and add_mono_wrt_lt lemmas and simplifying the result arithmetically , we can ??? it into ??? ???
- When ??? , we assume the induction hypothesis (N-1) ·A< (N-1) · B . Using ??? lt_to_le_rw we rewrite the hypothesis a<b and the induction hypothesis to A+1=B and (N-1) · A+1= (N-1) · B .
- When 1<N , we assume the induction hypothesis (N-1) ??? (N-1) · B . Using ??? lt_to_le_rw we rewrite the hypothesis a<b and the induction hypothesis to A+1=B and (N-1) · A+1= (N-1) · B .
- When 1<N , we assume the induction hypothesis (N-1) · A< (N-1) · B . Using lemma ??? we rewrite the hypothesis a<b and the induction hypothesis to A+1=B and (N-1) · A+1= (N-1) · B .
- In the step case , we have ??? , and (N-1) · A< (N-1) · B . Using ??? lt_to_le_rw , we know that A+1=B and (N-1) · A+1= (N-1) · B ,
- In the step case , we have 1<N , and (N-1) ??? (N-1) · B . Using ??? lt_to_le_rw , we know that A+1=B and (N-1) · A+1= (N-1) · B ,
- In the step case , we have 1<N , and (N-1) · A< (N-1) · B . Using lemma ??? , we know that A+1=B and (N-1) · A+1= (N-1) · B ,
- Since ??? ...0 and n: ...-1 , we ??? use div_3_to_1 to rewrite the ??? side of the equation a- ((-a)5(-n)) · n ??? - (-a-((-a)5(-n))· (-n)) .
- Since a: ??? and n: ...-1 , we can use ??? to ??? the ??? side of the equation a- ((-a)5(-n)) · n ??? - (-a-((-a)5(-n))· (-n)) .
- Since ??? ??? and n: ...-1 , we can use div_3_to_1 to rewrite the left side of the equation a- ((-a)5(-n)) · n ??? - (-a-((-a)5(-n))· (-n)) .
- then , we can use ??? ??? : ??? and get : a- ((-a)5(-n)) · n = - (-a-((-a)5(-n))· (-n))

- then , we can use lemma ??? : ??? and get : a- ((-a)5(-n)) · n ??? - (-a-((-a)5(-n))· (-n))
- by ??? ??? on the ??? div and some arithmetic a- ((-a)5(-n)) · n = - (-a-((-a)5(-n))· (-n))
- by ??? ??? on the leftmost ??? and some arithmetic a- ((-a)5(-n)) · n = - (-a-((-a)5(-n))· (-n))

References

Ahrenberg, L., Merkel, M., Hein, A. S., and Tiedemann, J. (2000). Evaluation of word alignment systems. In *Proceedings of the Second International Conference on Linguistic Resources and Evaluation (LREC-2000)*, volume III, pages 1255–1261.

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical machine translation: Final report. Technical report, Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Speech and Language Processing, Baltimore, MD. Available at `http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps`.

Barzilay, R. and Lee, L. (2002). Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Constable, R. L., Allen, S. F., Bromley, H. M., Cleaveland, W. R., Cremer, J. F., Harper, R. W., Howe, D. J., Knoblock, T. B., Mendler, N. P., Panangaden, P., Sasaki, J. T., and Smith, S. F. (1986). *Implementing Mathematics with the Nuprl Development System*. Prentice-Hall, NJ.

Coscoy, Y., Kahn, G., and Thery, L. (1995). Extracting text from proofs. In Dezani-Ciancaglini, M. and Plotkin, G., editors, *Proc. Second International Conference on Typed Lambda Calculi and Applications, Edinburgh, UK*, volume 902, pages 109–123.

Holland-Minkley, A. M., Barzilay, R., and Constable, R. L. (1999). Verbalization of high-level formal proofs. In *AAAI/IAAI*, pages 277–284.

Huang, X. and Fiedler, A. (1997a). Proof verbalization as an application of NLG. In *IJCAI (2)*, pages 965–972.

Huang, X. and Fiedler, A. (1997b). Proof verbalization in proverb. In *Proceedings of the First International Workshop on Proof Transformation and Presentation (PTP)*, pages 35–36.

Knight, K. (1999). A statistical mt tutorial workbook. Available at `http://www.clsp.jhu.edu/ws99/projects/mt/`. Prepared in connection with the JHU summer workshop.

Melamed, I. D. (1997). A word-to-word model of translational equivalence. In Cohen, P. R. and Wahlster, W., editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 490–497, Somerset, New Jersey. Association for Computational Linguistics.

Och, F., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proceedings EMNLP/WVLC*.

Och, F. J. and Ney, H. (2000a). A comparison of alignment models for statistical machine translation. In *COLING '00: The 18th Int. Conf. on Computational Linguistics*, pages 1086–1090, Saarbrücken, Germany.

Och, F. J. and Ney, H. (2000b). Improved statistical alignment models. In *ACL00: Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hongkong, China.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen.

Wang, Y.-Y. (1998). *Grammar Inference and Statistical Machine Translation*. PhD thesis, Carnegie Mellon University.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.