Toward Active Learning in Data Selection: Automatic Discovery of Language Features During Elicitation

Jonathan H. Clark, Robert Frederking, Lori Levin

Language Technologies Institute
Carnegie Mellon University
{jhclark,ref,lsl}@cs.cmu.edu

Abstract

Data Selection has emerged as a common issue in language technologies. We define Data Selection as the choosing of a subset of training data that is most effective for a given task. This paper describes *deductive feature detection*, one component of a data selection system for machine translation. Feature detection determines whether features such as tense, number, and person are expressed in a language. The database of the *The World Atlas of Language Structures* provides a gold standard against which to evaluate feature detection. The discovered features can be used as input to a Navigator, which uses active learning to determine which piece of language data is the most important to acquire next.

1. Introduction

In recent years, Data Selection has become a common issue in many areas of language technologies including speech recognition (Zhang and Rudnicky, 2006), speech synthesis (Black and Lenzo, 2001), and machine translation (Probst and Lavie, 2004). The benefits of data selection become evident at two stages. First, in gathering language data from humans, the development cost of language technologies is typically increased greatly. Second, in executing training algorithms, time (often on the order of days) is lost due to the mass quantities of data involved. However, much of this data is essentially redudant. Data Selection allows for gathering and/or using much less data while still retaining comparable levels of performance. While previous methods have performed data selection as a static step either before or after the collection of data from human sources, we use active learning to dynamically select data during data collection.

This paper describes *deductive feature detection*, which is one step in our active learning system. Feature detection answers answers questions such as "Does this language distinguish singular nouns from plural nouns?" It is employs active learning to characterize the current state when determining what data to select.

One might think that because some kind of reference grammar exists for almost every language, feature detection might be unnecessary. However, in the database of the *The World Atlas of Language Structures* (WALS) (Haspelmath et al., 2005), a compendium of the types of features we are trying to automatically detect, over 84% of the cells are blank. Haspelmath (Haspelmath, 2008; Haspelmath, 2007) notes that the data required to fill in most of these cells is not easily obtainable. Furthermore, reference grammars report on which features are *grammaticalized*, but many non-grammaticalized features are expressed. For example, future tense is not grammaticalized in Japanese, but may be expressed by circumlocutions such as *watashi wa gakoo ni iku yode desu* (lit. *"I have a plan to go to school."*) for *I will go to school*.

1.1. Corpus Design and Elicitation

The AVENUE Elicitation Corpus is a list of sentences associated with phrase structure trees and feature structures similar to the tectogrammatical layer in the Prague Treebank (Petr Sgall, 2004). These sentences are annotated with a head mapping and a ϕ mapping (Kaplan, 1995), which links nodes of the phrase structure tree to nodes of the feature structure, thereby providing a mapping from words to the features that they express (Figure 1). The corpus is organized into *minimal pairs* each bringing out a certain language feature. For example, the corpus includes sentences

such as *Maria bakes cookies*, *Maria baked cookies*, *She bakes cookies*, etc. In this way, we maximize the coverage of our corpus with regard to language features (but unfortunately not vocabulary).

During elicitation, a relatively naive bilingual person translates the sentences and aligns the words via a GUI. The result of elicitation is a parallel, word-aligned corpus with the mappings described above. This corpus can then be used for learning correspondences between semantic structures in the two languages. In the case of AVENUE, this involves the automatic extraction of MT transfer rules (Carbonell et al., 2002). Though we typically only elicit around 4,000 sentences, we hypothesize that we can make this small amount of data more valuable by using active learning to select the sentences from which the most effective translation rules can be learned (Probst and Levin, 2002).

Note that not all language features that we might want to elicit manifest themselves in English. In these cases, we provide the bilingual person with a context field to elicit the proper meaning. For example, if we wanted to elicit dual number (as would be necessary for Arabic), we might provide the sentence *Men bake cookies* to be translated alongside the context of *Translate this as if there are two men*.

1.2. Corpus Navigation

Early work used a static set of source language sentences for every target language (Alvarez et al., 2006), but not all data is relevant for every language. Further, the space of possible features to be explored is simply too large (millions of sentences, according to our current feature specification). Since feature detection relies on having a set of input sentences that exemplify the necessary language features, it is important that we select an appropriate subset of all possible combinations of language features to feed to Feature Detection.

In Corpus Navigation, we apply feature detection to the set of sentences that have already been acquired from the bilingual speaker and then use the resulting knowledge to choose the most valuable sentences to be elicited next. There are two inputs to Corpus Navigation: (1) the space of grammatemes (features that might affect syntax and morphology) and how they might interact and (2) the presence or absence of these features discovered so far (the *current state*). The Corpus Navigator returns as output a recommentation for the most useful piece of information to ask next.

At this stage, Greenbergian Typological Universals (Greenberg, 1963) can be used to augment the knowledge in our *current state*. For example, if a language does not distinguish singular nouns from plural nouns, then we know the language will not grammaticalize dual number, and we can direct our search through the corpus accordingly. Recently, methods have been developed for the automatic discovery of Typological Univerals (Daumé and Campbell, 2007), making a large number of univerals available in digital format for tasks such as Corpus Navigation.

Finally, we point out that this paper describes one component of a Corpus Navigation system (Figure 2): the *deductive feature detection* component, which detects how features are expressed by using a set of rules that take annotations of one feature set as input and fire implications of a different feature set as output. Other components could be added to the Navigator such as *inductive* feature detection in which feature expression is analyzed over feature with which the corpus is annotated. Note that such a component would not be able to interact with the WALS database unless the corpus were tagged with WALS features. In addition to the Typological Universals described above, Figure 2 shows a structural analysis component, which could navigate the system toward discovering more interesting constituent structure patterns, as well as, a morphology analyzer, which could guide navigation toward sentences that would be helpful in automatically inducing the morphology of a languages. The Navigator itself will rely on a search heuristic informed by a linguistic knowledge base of genetic and areal typology.

1.3. Feature Detection

The first step in Corpus Navigation is to define the *current state*. For this, we need to detect language features by comparing minimal pairs of sentences. For example, with the above examples, we might detect that Spanish marks past tense by comparing the sentence pairs { *[Maria bakes cookies / Maria hornea galletas]*, *[Maria*

context:	Maria bakes cookies often or habitually.
srcsent:	Maria bakes cookies .
tgtsent:	Maria hornea galletas .
aligned:	((1,1),(2,2),(3,3),(4,4))
fstruct:	[f1]([f2](actor ((gender f)(anim human))) [f3](undergoer ((person 3))) (tense pres))
cstruct:	[n1](S1 [n2](S [n3](NP [n4](NNP Maria))
	[n5](VP [n6](VBZ bakes) [n7](NP [n8](NNS cookies)))))
phimap:	phi(n1)=f1; phi(n3)=f2; phi(n7)=f3;
headmap:	h(n1)=n2; h(n2)=n5; h(n3)=n4; h(n4)=n4; h(n5)=n6; h(n6)=n6; h(n7)=n8; h(n8)=n8;

Figure 1: Here, we see a post-elicitation example entry from the elicitation corpus. Note that the target language sentence and alignments were added via a bilingual person's interactions with a GUI.

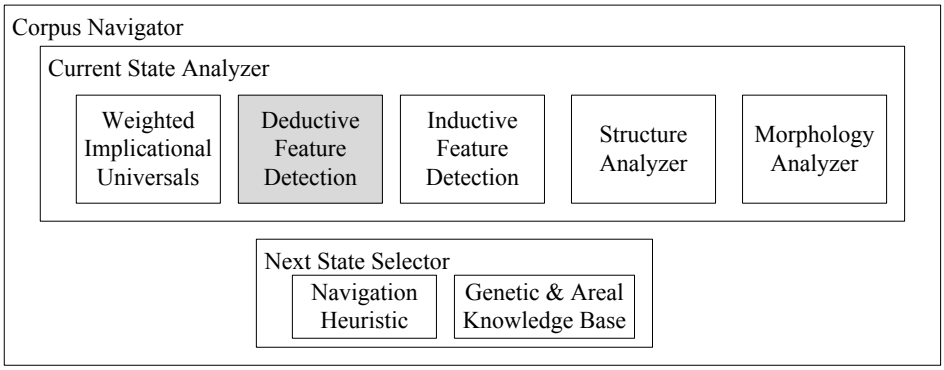

Figure 2: This figure shows the various components that might eventually be involved in a Corpus Navigator. This paper describes the *deductive feature detection* component, which detects how features are expressed by using a set of rules that take annotations of one feature set as input and fire implications of a different feature set as output.

baked cookies / Maria horneó galletas]}, and the words { *[bakes / hornea]*, *[baked / horneó]*}. The remainder of the paper describes the implementation and evaluation of feature detection.

2. Data

The input for experiments in this paper is a Spanish-English elcitation corpus of 50 sentences. We selected 21 features, shown in Table 1, from WALS for this experiment. The output of feature detection is a value for Spanish for each of the features. For example, the value for `Occurrence of Nominal Plurality` should be `Yes`.

The Linguistic Data Consortium now provides Language Packs for Less Commonly Taught Languages (LCTL) for Thai, Bengali, Urdu, each including 3100 sentences from our elicitation corpus (Levin et al., 2006). After alignments have been obtained for these corpora, they will serve as a primary area of application for feature detection.

3. Implementation

We represent the mapping between example sentences and language features by a set of LISP-like production rules, which can be reused for arbitrary language pairs (Figure 3). These are read by a Java interpreter, which takes as input the rules and the sentences elicited so far.

Each production starts with a sentence pattern, which matches a piece of the feature structure such as `(aspect perfective)` and `(aspect imperfective)` in Figure 3. Each production rule can have multiple conditions each of whichproposes a feature value. The rule in Figure 3 has two conditions: the head verbs of two sentences are identical or they are different. If

the head verbs are identical, the proposed value of `Perfective/Imperfective Aspect` is `Grammatical Marking`.

Because languages have irregularities and subregularities, it is common for several different feature values to be proposed when the rules apply to several sentences. For example, in a language that optionally marks plurals, both `Marks Plural` and `Does Not Mark Plural` may be proposed. This ambiguity is resolved by assigning a λ value (weight) to each condition. For a set of conditions C in a rule with each $c \in C$ being asserted n_c times, the condition (feature value) v asserted for a language is given by $v = \text{argmax}_{c \in C} \lambda_c n_c$.

In evaluating each condition, the system also checks for degree of lexical overlap, which may be specified in the rule. This helps ensure that each comparison is really being done on a minimal pair. For each successful comparison, the system cites the evidence that led to that assertation including the matching sentences and their matching constituents.

4. Results

We evaluated the output of the system by calculating the precision, recall, and F1 scores of the output compared to the values recorded in WALS. As a baseline measure, we calculated the scores that would be obtained by choosing the most common value for each feature.

The results of the experiment are presented in Table 2. Since every rule fired exactly one final feature value, the precision, recall and F1 scores are equivalent. As error analysis, we present two of the features that did not fire the values recorded in WALS: `Order of Subject and Verb`, which fired `SV` instead of the expected `No dominant order` and `Verbal Person Marking`, which fired `Only the A argument` when `Both the A and P arguments` was expected. In both cases, the values fired by the system are reasonable given the input sentences. Given a larger set of input sentences, it is likely that these values could have also been detected correctly. In the future, we would like to explore methods for detecting when we have "enough" data.

In terms of the Corpus Navigation Application (Section 1.2.), these results are promising. Given that we can now detect with good accuracy the proper values of many WALS features given a set of annotated sentences, we can now query the Typological Universals discovered in Daumé and Campbell (2007) to determine which other features might be expressed. This, in turn, informs our active learning feedback loop that we should explore sentences that might display the predicted

Feature Name
Gender Distinctions in Independent Personal Pronouns
Nominal and Locational Predication
Occurrence of Nominal Plurality
Order of Adjective and Noun
Order of Genitive and Noun
Order of Numeral and Noun
Order of Subject, Object and Verb
Order of Subject and Verb
Order of Object and Verb
Perfective/Imperfective Aspect
Politeness Distinctions in Pronouns
Position of Interrogative Phrases in Content Questions
Position of Pronominal Possessive Affixes
Position of Tense-Aspect Affixes
Inclusive/Exclusive Distinction in Independent Pronouns
Inclusive/Exclusive Distinction in Verbal Inflection
Semantic Distinctions of Evidentiality
The Future Tense
Verbal Person Marking
'Want' Complement Subjects
Zero Copula for Predicate Nominals

Table 1: The 21 features chosen from WALS to evaluate feature detection.

	Precision	Recall	F1
Experimental	85.71%	85.71%	85.71%
Baseline	57.14%	57.14%	57.14%

Table 2: Evalution of the feature detection system on selected features from the WALS database. The baseline was calculated by choosing the most frequent feature value across the languages surveyed in WALS.

```
# Perfective/Imperfective Aspect
(rule (variables (PERFECTIVE ((aspect perfective)))
                 (IMPERFECTIVE ((aspect habitual)
                                (aspect progressive))))
      (sentences(A (PERFECTIVE))
                (B (IMPERFECTIVE)))
      (overlap   (default))
      (if 0.9    (different (target-lex-uhead (fnode (A))) (target-lex-uhead (fnode (B))))
                 (then (WALS "Perfective/Imperfective Aspect" "Grammatical marking")))
      (if 0.1    (same (target-lex-uhead (fnode (A))) (target-lex-uhead (fnode (B))))
                 (then (WALS "Perfective/Imperfective Aspect" "No grammatical marking"))))
```

Figure 3: This figure shows an example rule from the production system that is used for feature detection. The first "if" statement denotes that if the ultimate head of target lexicons mapped to the grammatical aspect being perfective are different than those mapped to imperfective, the value "Grammatical marking" should be fired. Since $\Sigma\lambda = 1.0$ in this case, if 90% of the evidence in this rule suggest (propose) this feature value, then it is asserted as the correct value.

features. Since this process will be used as a search heuristic here, we believe that the reported accuracies are reasonable for this task.

5. Language Resources

The result of Corpus Navigation is a resource dense with the "right" features, those for which the target language makes distinctions. This corpus is also highly structured in that each language feature is linked with sentences illustrating that feature. It is also valuable in that it is word-aligned and each sentence is associated with a a feature structure that expresses the meaning of the sentence. Such a resource can be valuable for studying MT divergences (different ways of expressing the same meaning) and for studying constructions (partially non-compositional form-meaning mappings). In effect, it is a parallel treebank annotated with an interlingua of grammar.

6. Applications

The process of Feature Detection enables us to detect language features and how they are expressed. In learning these morphosyntactic language features using the Elicitation Corpus, we can learn most of a language's syntax and structure without a large parallel corpus. We can then leverage alternative language resources by using the hybrid MT System AVENUE, which we say is "omnivorous" in that it can make use of many different kinds of language resources during system construction. This enables us to use other lexical resources such as bilingual dictionaries and monolingual corpora to address the issue of lexical coverage. Further, due to the unstructured nature of parallel corpora, there are some tasks (such as automatically annotating closed-classe morphemes with their corresponding language features) that would be nearly impossible with parallel data. By detecting language features from an elicitation corpus, we can perform tasks that would otherwise be too difficult, and we leave open our options for how to deal with lexical coverage.

7. Conclusion

Having developed a system that can automatically discover language features during the elicitation process, we now look toward the development of a Corpus Navigator. These tools will aid in the economical creation of parallel corpora that are dense with the characteristic features of each target language.

8. Acknowledgements

Thanks to our AVENUE and LTI colleages for their support, to Hal Daumé for providing a programatically readable version of the WALS database, to José P. González for verifying the

Spanish-English test corpus, and to Keisuke Kamataki for the Japanese example. Our appreciation also goes to Martin Haspelmath for his helpful insights into the nature of the gaps in the WALS database. Finally, we are grateful for the helpful comments of three anonymous reviewers. This work was supported by US National Science Foundation Grant Number 0713-292.

9. References

Alison Alvarez, Lori Levin, Robert Frederking, Simon Fung, Donna Gates, and Jeff Good. 2006. The MILE corpus for less commonly taught languages. In *HLT-NAACL*, New York, New York, June.

Alan Black and Kevin Lenzo. 2001. Optimal data selection for unit selection synthesis. In *ISCA, 4th Speech Synthesis Workshop*, Scotland.

Jaime Carbonell, Kathrina Probst, Erik Peterson, Christian Monson, Alon Lavie, Ralf Brown, and Lori Levin. 2002. Automatic rule learning for resource limited MT. In *Association for Machine Translation in the Americas (AMTA)*, October.

Hal Daumé and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Joseph Greenberg. 1963. *Universals of Languages*. MIT Press, Cambridge.

Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie. 2005. *The World Atlas of Language Structures*. Oxford University Press, Oxford.

Martin Haspelmath. 2007. personal communication.

Martin Haspelmath. 2008. The typological database of the *World Atlas of Language Structures*. In Martin Everaert and Simon Musgrave, editors, *Typological databases*. Mouton de Gruyter, Berlin.

Ronald Kaplan. 1995. The formal architecture of lexical functional grammar. In Mary Dalrymple, Ronald Kaplan, J. Maxwell, and A. Zaenen, editors, *Formal Issues in Lexical Functional Grammar*. CSLI Publications.

Lori Levin, Jeff Good, Alison Alvarez, and Robert Frederking. 2006. Parallel reverse treebanks for the discovery of morpho-syntactic markings. In *Proceedings of Treebanks and Linguistic Theory*, Prague.

Eva Hajicová Petr Sgall, Jarmila Panevová. 2004. Deep syntactic annotation: Tectogrammatical representation and beyond. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation, Association for Computational Linguistics*, Boston.

Katharina Probst and Alon Lavie. 2004. A structurally diverse minimal corpus for eliciting structural mappings between languages. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, Washington, DC.

Katharina Probst and Lori Levin. 2002. Challenges in automated elicitation of a controlled bilingual corpus. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-02)*.

Rong Zhang and Alexander I. Rudnicky. 2006. A new data selection approach for semi-supervised acoustic modeling. In *Acoustics, Speech and Signal Processing (ICASSP)*.