Expanding lexicons by inducing paradigms and validating attested forms

Gregory Grefenstette, Yan Qu, David A. Evans

Clairvoyance Corporation
5001 Baum Blvd, Pittsburgh, PA 15213, USA
{g.grefenstette, y.qu, d.evans}@clairvoyancecorp.com

Abstract

One of the bottlenecks in Natural Language Processing for a given language is creating a lexicon that covers the language. The morphological lexicon provides two important pieces of information for NLP applications: 1) the normalization of a word, its lemmatization, which allows the application to recognize two variants of the same word; and 2) the part-of-speech roles that the word can play, which allows the application to parse the text, creating relations between the words in a text. Many NLP applications, e.g. Information Retrieval, Classification, Terminology Extraction, etc., depend upon the normalization and parsing information found in lexicons. When words are not present in these lexicons, it is difficult to predict what their proper lemmatizations and parts-of-speech are. In this paper we present a technique for updating a lexicon given an unknown word via induction of paradigms from an existing, but incomplete, lexicon and validation of the paradigm using corpus evidence.

1. Introduction

All computer programs that process natural language must use morphological lexicons to recognize when two different word forms refer to the same word (e.g. *thought* as a form of *think*) and/or to know what grammatical role a word can play (noun, verb, etc.). One problem with these lexicons is the cost and delay of manually updating them. Through a combination of web spidering and language guessers, it is possible to continually find new words to add to these lexicons, but the problem remains of deciding which paradigms apply to which newly found words. We present here a method for inducing paradigms from simple lists of words, normalization and part-of-speech, and then show how these are applied to new words.

2. Related Work

The problem of handling words not in the lexicon has long been a concern of Natural language Processing (NLP) systems. In NLP of written text, part-of-speech tagging systems have treated this problem in a number of ways. One popular method is to create a list of word suffixes from known lexicon words and to use the longest matching suffix of an unknown word to assign a most common part-of-speech tag (Kupiec, 1992; Meteer et al., 1991). Another method tries to map unknown words to known words in the lexicon by adding or deleting affixes (Mikheev, 1994; Black et al., 1991). These methods provide a part-of-speech tag for the individual unknown word but provide neither a lemmatization of the word nor a prediction as to what other words are variants of the same word. Some work has been performed on creating word families by automatically inducing stemming rules though not necessarily part-of-speech information for a language from a large quantity of tokenized text. (Oard et al., 2001; Goldsmith, 2001). Our work here unites these three methods.

We use both a lexicon and a large corpus. The lexicon yields paradigms for the language. Any number of these paradigms can be applied to an unknown word. The large corpus provides a word list that allows us to choose which paradigm is best attested for the unknown word. The results of our work are new, full lexical entries with both part-of-speech tags and normalized lemmatizations.

3. Existing Lexicons

For many written languages, linguists have already developed lexicons. In what follows, we suppose that one has access to a large but incomplete lexicon of word forms which contains, for each surface form (i.e. as the word appears in text) a lemmatization of that surface form and part-of-speech information for that surface form. We use in these experiments the lexicons that were developed by the Multext project (Ide & Veronis, 1994) for English, Spanish, Italian, German and French. An example of entries in these dictionaries is the following from the Spanish lexicon:

abogadas	abogado	Ncfp-
abogado	abogado	Ncms-
abogados	abogado	Ncmp-
abominable	abominable	Afp.s-
abominablemente	abominable	Rg
abominables	abominable	Afp.p-

in which the surface form appears in the first column, the lemma in the second column, and part-of-speech and morphological information in the third column. The first letter of this last column gives the major part of speech category (N=noun, A=adjective, R=adverb). The interpretation of the second letter depends on the first letter. 'Ncfp' means 'Noun, common, feminine, plural', whereas 'Afp.s' means 'Adjective, qualitative, positive, masculine/feminine, singular.' For more information on the creation of these lexicons, see http://www.lpl.univ-aix.fr/projects/multext.

The techniques developed below can directly apply to any lexicon expressible in this format:
surface form -- lemma -- tags.

3.1. Unknown Words

As fodder for updating an incomplete lexicon, we use a list of unknown words found in texts in that language. It is possible to grab a large quantity of text in a given language (Ghani et al., 2001) using language identification (Grefenstette, 1995) techniques. As a shortcut, instead of building a large corpus, we decided to use a large list of words that was developed for a language specific spell checker. The large list of words used for the GNU *Ispell* (for international spell checker) can easily be derived from sources found at http://ficus-www.cs.ucla.edu/project-members/geoff/ispell-dictionaries.html. The Ispell lists are intended to be as complete as possible, even to the point of overgeneration, so that any correctly spelled word belonging to the language would not be flagged as an unknown misspelled word.

Comparing this list of words and the Multext lexicon we wish to expand in our experiments, the Multext Spanish lexicon contains 474,159 unique surface forms, whereas the Ispell Spanish wordlist contains 730,003 unique surface forms, 483,517 of which are not found in the Multext Spanish lexicon.[1] It is these half a million unknown words that we wish to include in the lexicon automatically using the morphological paradigms derived from the existing lexicon.

We will also use both the words in this Ispell list and the words the Spanish Multext lexicon as attested words to decide which induced paradigms to apply.

4. Inducing Morphological Paradigms

Morphological paradigms can be seen as a set of patterns describing how a given lemma generates surface forms. For example, one paradigm applicable to the example lemma 'abogado' given above could be described as

> *given a noun ending in –o, form the masculine plural by adding an –s, form the feminine singular by removing the –o and adding an –a, and the feminine plural by removing –o and adding –as.*

In this paradigm, we only consider suffixes to be deleted and/or added. We will call the letters appearing before the suffixes *context* for the paradigm. The above paradigm for `abogado` is a 0-context paradigm since no account is taken of the letters preceding the suffixes.

One might restrict the paradigm to only those words ending in –do (1 letter of context, called 1-context below), or to those ending in –ado (2 letters context, called 2-context below), etc.

We created all the 0-context paradigms from the Multext Spanish lexicon by collecting all the surface forms associated with each lemma and stripping off letters from the lemma and the surface form until the stems matched. This resulted in 416 distinct 0-context paradigms, 833 more restrictive 1-context paradigms, 1759 2-context paradigms, and 3607 unique 3-context paradigms.

4.1. Applying Paradigms to Unknown Words and Attested Validation

Given an unknown word (e.g. from the Ispell word list), we wish to assign it to its best paradigm[2]. This paradigm is unknown. Each element of the paradigms that we tested, as we have shown above in the example, indicates what suffix to remove and what suffix to add in order to find the lemma. Given an unknown word that matches the suffix of a pattern, we generate a candidate lemma. This candidate lemma, in turn, when applied to the other patterns in the paradigm, generates a number of new candidate surface forms. In the ideal case, all the candidate surface forms generated by applying the right paradigm to the right lemma would correspond to attested words found in a complete word list, approximated here by the combined list of words from Ispell and from the Multext lexicon[3]. In the normal case, certain paradigms will undergenerate and others will overgenerate.

4.2. Finding the Best Paradigms

To find which paradigm induction and application method gives the best results, we proceed as follows. Over 100 runs, we hold out a part of the Multext Spanish lexicon. This held out set will be used as a test set and the actual surface forms associated with these words in the Multext lexicon become the gold standard for the run. In Table 2, we present the results of the following experiments.

Over 100 runs, for each run,

- We extracted 100 randomly chosen lines from the Spanish Multext lexicon. We limited the extracted words to noun, verbs and adjectives, since these are the open class categories (in the Spanish Multext, there are 12970 open class lemmas and 267 closed-class lemmas) and we suppose that the basic lexicon that we are looking to expand already contains the closed class words. For adverbs derived from open class words see the next point. We call this set of words HELDOUT for this run.
- For each extracted word in HELDOUT, we retained the lemma and then removed all surface forms associated with that lemma from the Multext lexicon. Note that this removal also removed any derived adverbs since the Multext lexicon maps these adverbs to noun, verb or adjective forms. All removed lexical entries are put aside in a set we call GOLDSTANDARD. The reduced lexicon (original lexicon less the extracted entries) is called the REDUCED LEXICON.

[1] The Multext lexicon also contained 227685 surface forms not present in the Ispell list. E.g., bemolmente, benedictinamente, beneficiosamente, benitamente, bestiamente, bicolormente, bidireccionalmente, …

[2] Actually, we sometimes choose more than one paradigm for a word. A word may have a complete verb paradigm and a complete noun paradigm, for example. These complete, attested paradigms are called 100% matches in the text and in Table 2.

[3] One could also use a word list generated from a very large corpus, or use a WWW portal, such as Altavista, that would be polled to attest the existence of a word form.

- Using the REDUCED LEXICON, we generated three different paradigm sets. For each lemma remaining in the REDUCED LEXICON, we extract the set of all surface forms that correspond to it, and for that lemma we create a paradigm using different context lengths:
  - 0-context paradigms: match each lemma to each surface form, starting from the leftmost letters until the surface form and the lemma diverge. Retain the divergent letters and the parts-of-speech of the surface form as one pattern in this lemma's paradigm.
  - 1-context paradigms: same as above but retain the last matching letters as part of the pattern.
  - 2-context paradigms: retain the last two matching letters in the patterns.
  - 3-context paradigms: retaining the last three letters before suffix removal.

  Table 1 gives the most frequent 0-context, 1-context, 2-context and 3-context paradigms derived from the Multext Spanish lexicon.
- For each context length, we retained either the top 100 most frequent paradigms, or the top 200, or the top 300, or all the paradigms.
- For each of the 16 SET OF PARADIGMS (that is, using 0 to 3 letters of context and the top 100, 200, 300 or entire set of paradigms extracted from the REDUCED LEXICON), we applied the extracted paradigms to the testing words in HELDOUT.
- For each paradigm in a SET OF PARADIGMS and each word in HELDOUT, we examined whether any pattern in paradigm was applicable to the word (i.e., did the ending of the word match the suffix to be added). If the pattern could be applied then the CANDIDATE LEMMA was created and all the surface forms corresponding to that paradigm were created, making a set of CANDIDATE SURFACE FORMS.
- Then we calculated, for that word, for each paradigm, how many of the CANDIDATE SURFACE FORMS were attested in the list of Ispell Spanish forms. This number was used to score the paradigm. Each paradigm had two scores: the number of attested surface forms (ATTESTED) that the paradigm generated, and the PERCENTAGE of paradigm patterns that generated an attested surface form.
  - Here is an example. Suppose that we have in HELDOUT the imaginary word *zoza* and that we are applying the first paradigm shown in Table 1. The second pattern in this paradigm produces the CANDIDATE LEMMA *zozo*, and the whole paradigm predicts that CANDIDATE SURFACE FORM *zozo*, *zoza*, *zozos* and *zozamente* exist. Now suppose that in the Ispell list we find only *zozo*, *zoza*, *zozos*, then this paradigm has a ATTESTED SCORE of 3, and a PERCENTAGE score of 3/4 or 75%.
- After testing all the paradigms in a given word in the HOLDOUT set, the paradigm having the best score for a given word was accepted for that word and all the surface words corresponding to that word and that paradigm were generated. We tested three ways of calculating the best score:
  - always taking the paradigms that had the most attested candidate surface forms (columns MOST ATTESTED in Table 2);
  - always taking the paradigms with the most attested surface forms, plus any other paradigm that had a 100% match of all its candidate surface forms (MOST ATTESTED PLUS ALL 100%);
  - always taking the paradigms with the best percentage of attested surface form matches, and if this is percentage is 1000%, taking also any other paradigm that had a 100% match of all its candidate surface forms (BEST PERCENT PLUS 100% in Table 2).

5. Results

We performed 100 runs with each of the 16 configurations of using 0 to 3 letters of pre-suffix context, and only the most frequent to all of the paradigms generated. For each run, we used one of three ways of calculating the best paradigm(s) to choose for each word in the HELDOUT set, and verified the lexicon entries generated using the best paradigms for each word (i.e. the paradigms that generated the most overlap with the raw list of words in the Ispell list) against the original Multext entries for these words which were stored apart in GOLDSTANDARD. We calculated the precision and recall of all new entries against this GOLD STANDARD. Precision gives the percentage of generated entries that were found in GOLDSTANDARD, and recall gives the percentage of GOLDSTANDARD entries that were found in the list of generated entries. In the ideal case, these lists would be identical and then precision and recall would both be 100%. The results of averaging the 100 runs for every configuration of paradigm extraction and number of paradigms used are given in Table 2.

For example, in this table, we see that using 1 letter of context before suffix removal, and using the 200 most frequent paradigms extracted from the REDUCED LEXICONS over 100 runs, and using the paradigm that gave the highest percentage of overlap with the GOLD STANDARD gave us an average recall of 90% of the GOLD STANDARD entries with an average precision of 68%, i.e., 68% of the lexical entries generated by this configuration actually appeared in the GOLD STANDARD for each run.

In general, Table 2 shows that precision improves when more context is used, and recall improves, for a given level of context, when more examples are used, but that using less context improves recall the most.

We see a trade-off between recall and precision. For example, using paradigms built using zero letters of suffix context, and including all the paradigms extracted from the Multext lexicon, we reach 97-98% recall of correct lexical entries (i.e. entries that are found in the GOLDSTANDARD), but these correct entries are accompanied by over-generation of many spurious entries, so that the precision of all the entries produced is only 34-39%.

We also present in this table a measure that combines Precision and Recall, called the F-measure. Here the F-measure is calculated by adding the Precision to the Recall and dividing by 2. This measure gives a single number that balances the influence of precision and recall. If the lexical entries generated by the system were to be manually verified by a lexicographer then one might want to favor recall, since it would be easy to present newly generated lexical entries for acceptance or rejection. In a completely automated system, one might use the F-measure to select a configuration that is pretty good both in terms of recall and in terms of precision.

For an automated system, we would prefer using the configuration consisting of using 1 letter of context and using the 100 most frequent paradigms as the best combination, giving an F value of 76 with a minimum number of paradigms to store and apply.

An example of a paradigm that is incorrectly recognized in a configuration with a high F-score is the following. With 1 letter of pre-suffix context, taking the top 100 paradigms, and taking as the best paradigm the one that has the most attested forms in the Ispell list, in run 34, we find the word *falda (skirt)* which is erroneously recognized as a verb (like *escaldar* and *respaldar*) and which is fully conjugated. One the other hand, the HELDOUT list for this run contained the word *pecado (sin)*, which provoked the generation of the verb form *pecar (to sin)*. Since only the noun *pecado* was present in the Multext GOLDSTANDARD, all the verb forms were considered noise. Another common error was confusion between adjectives and nouns. For example, the HELDOUT word *psicótico* was considered as a noun and as an adjective (like *académico)* but the GOLDSTANDARD Multext lexicon only contains the adjectival entries so all the noun readings are considered as noise.

6. Conclusions

We assume in these experiments that all irregular paradigms are initially present in the lexicon, a presumption that may hold in a case like the Multext lexicons which were built by computational linguists. Applying our preferred configuration (1 letter context and using the top 100 most frequent paradigms extracted from the Multext lexicon) creates new MULTEXT lexical entries (lemmatization and part of speech tags) in a matter of minutes for all the half-million new words present in the Ispell list. If we limit the new entries to words whose forms appear at least once on the WWW (using a portal such as Altavista to obtain these counts) then we still have 349875 new entries to add to the Multext Spanish lexicon. It will also be possible to continually enrich the lexicon by applying the same technique to new words found on web pages known to be in the specified language (by daily downloading of newspapers, for example).

7. References

Black, A., J. van de Plassche, and B. Williams. 1991. *Analysis of Unknown words through Morphological Decomposition*. In Proceedings of the 5th Conference of the EACL, Volume 1, 101-106.

Goldsmith, J. 1998. *Unsupervised learning of the morphology of a natural language*. Unpublished Manuscript, available at http://humanities.uchicago.edu/ faculty/ goldsmith/index.html.

Ide, N. and J. Veronis. 1994. *Multext: Multilingual text tools and corpora*. In 15th International conference on computational linguistics (COLING), Kyoto, Japan, pp. 588—592.

Ghani, Rayid, Rosie Jones, and Dunja Mladenic. 2001. *Using the Web to Create Minority Language Corpora*. In Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM).

Grefenstette, G. 1995. *Comparing two language identification schemes*. In Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95), vol 1, pp. 263-268, Rome, Dec 11-13, ISBN 88-7975-159-X.

Koskenniemi, K. 1991. *A Discovery Procedure for Two-Level Phonology*. In L. Cignoni and C. Peters (editors), Computational Lexicology and Lexicography: A Special Issue Dedicated to Bernard Quemada, 1991, 451-465**.**

Kupiec, J. 1992. *Robust part-of-speech tagging using a hidden Markov model*. Computer Speech and Language, 6:225--243.

Mikheev, Andrei. 1997. *Automatic rule induction for unknown-word guessing*. Computational Linguistics, 23(3):405--423.

Meteer, M., R. Schwartz, and R. Weischedel. 1991. "*POST: Using probabilities in language processing*." In Proceedings of the 12th International Joint Conference on Artificial Intelligence, Sydney, Australia, 960—965.

Oard, D.W., G.-A. Levow, and C. I. Cabezas. 2001. *CLEF experiments at Maryland: Statistical stemming and backoff translation*. In C. Peters, editor, Proceedings of the First Cross-Language Evaluation Forum.

Most common 0-context paradigm,
corresponding to 2020 lemmas

Context	Delete suffix	Add suffix	New Tag
<null>	o	amente	Afpfp-
<null>	o	a	Afpfs-
<null>	<null>	s	Afpmp
<null>	<null>	<null>	Afpms-

Most common 1-context paradigm,
corresponding to 1469 lemmas

Context	Delete	Add	New Tag
a	<null>	<null>	Ncfs-
a	<null>	s	Ncfp -

Most common 2-context paradigm,
corresponding to 759 lemmas

Context	Delete	Add	New Tag
ci	ón	ones	Ncfp -
ci	ón	ón	Ncfs -

Most common 3-context paradigm,
corresponding to 543 lemmas

Context	Delete	Add	New Tag
aci	ón	ones	Ncfp -
aci	ón	ón	Ncfs-

Table 1. Most common paradigms extracted from the Spanish Multext lexicon, retaining zero, one, two or three letters of context before the suffixes to be deleted or added.

		Precision			Recall			F-Measure		
Context/ number paradigm used		Best percent plus all 100%	Most attested plus all 100%	Only Most attested	Best percent plus all 100%	Most attested plus all 100%	Only most attested	Best percent plus all 100%	Most attested plus all 100%	Only Most attested
0	100	43	39	27	**95**	**95**	41	59	55	33
	200	42	38	58	94	**98**	**95**	58	54	72
	300	39	34	56	**95**	**98**	94	55	51	70
	all	39	34	56	**97**	**98**	94	55	50	70
1	100	72	65	67	83	92	90	**76**	**76**	**76**
	200	68	62	66	90	**95**	93	**77**	74	**76**
	300	63	56	59	91	**95**	87	74	70	70
	all	55	50	61	**95**	**97**	91	69	66	73
2	100	75	72	74	43	58	57	54	64	64
	200	74	69	70	65	75	74	69	71	72
	300	74	68	69	73	82	81	73	74	74
	all	68	62	64	90	**95**	92	**77**	74	**75**
3	100	**85**	80	81	22	23	23	34	35	35
	200	**83**	77	78	35	36	35	49	49	48
	300	**83**	77	78	42	44	43	56	55	55
	all	74	67	68	82	88	87	**78**	**76**	**76**

Table 1 Precision, recall and F-measure results for combinations of pre-suffix context retained, and number of paradigms retained from training. Using more paradigms gives better recall but reduces precision. Our preferred configuration is retaining the 100 most frequently applicable paradigms, using 1 letter of context before the suffixes to be deleted and/or added in the paradigm patterns, and using the paradigms that gives the most attested number of forms (plus any full match paradigms). This configuration has an F-score of 76, a recall of 92%, and a precision of 62% (i.e. it overgenerates lexical entries but gets most of the good entries).