Automatic inference of word meaning using phonosemantic patterns[*]

Maria Ruiz-Casado, Enrique Alfonseca and Pablo Castells
Computer Science Department
Universidad Autonoma de Madrid
{Maria.Ruiz,Enrique.Alfonseca,Pablo.Castells}@uam.es

Abstract

This paper describes a new algorithm for discovering automatically possible meanings of unknown words from their phonetic features. It has been tested to obtain the semantics of a few invented verbs, using Levin's verb classes.

1 Introduction

The generalised availability of semantic metadata that describe (annotate) web contents has been identified as a fundamental issue for the semantic web to reach its critical mass. This is obviously an out-of-reach endeavour to be done manually, given the current size of the web. Therefore, automatic knowledge acquisition techniques are indispensable for automatic or semi-automatic ontology population and document annotation. This is usually done with the aid of heterogeneous techniques, including dictionaries, linguistic analyses, Information Extraction, statistical models and clustering. In this paper we propose a new algorithm which, based solely on phonetic features, tries to discover semantic connotations of unknown words.

Phonosemantics is the field of Linguistics that studies the relation between the meaning of the words and the way they are pronounced. Linguists and philosophers have tried for centuries to ascertain whether phonemes can convey meaning. These phonemes are usually called phonestemes. If this is the case, the meaning of a word should be influenced by the connotations of its constitutive phonemes, and this could be used for several applications, including automatic discovery of the meaning of unknown words for ontology population. In this field, efforts have focused on finding phonological patterns typifying a certain semantic. According to [Genette, 1976], already in 1653, [Wallis, 1653] publishes a list of phonestemes in English language. For illustration, he notes that

- */wr/* shows obliquity or twisting: *wry, wrong, wreck* and *wrist*.
- */br/* points to a breach, violent and generally loud splitting apart: *break, breach, brook*.

These efforts can be found in other languages likewise. For instance, [Marcos-Marín, 2001] finds the following two phonosemantic patterns for the Spanish language:

- *.* o .* o*, used in an adjective, adds to the word meaning the denotation of defect, physical or psychical limitation, e.g. *bronco, sordo, cojo, flojo*.
- *.* er .* o* means scorn, disrespect, e.g. *cerdo, lerdo, terco*.

[Magnus, 2001] describes several experiments regarding the classification of phonestemes and their clustering in semantic classes, and ends up with the conclusion that there is a relevant relation between semantics and phonetics, and that every word containing a given phoneme has some element of meaning that is lacking in words not containing that phoneme. She also concludes that the extent to which a word follows a phonosemantic rule is dependant on the degree of generality of the meaning of the word. That is to say: words that have a very concrete reference, with none or few exact synonyms, display less phonetic patterns related to their meaning.

2 Automatic phonosemantic analysis of unknown verbs

```
Initialise:
1. Obtain the phonetic transcription of all the words in
      the dictionary with no more than three syllables.

Get classes(word $w$):
1. Let $C$ = empty list.
2. Obtain $p$, the phonetic transcription of $w$.
3. Let $N$ = number of phonemes in $p$.
4. For $n = 0$ to $N - 2$,
      4.1. For every pattern $x$, calculated as any possible
            choice of removing $n$ phonemes from $p$,
            4.1.1. If there is not any class containing $x$, continue
            4.1.2. Locate $c$, Levin's verb class with more verbs
                  containing the phonetic pattern $x$ (if any).
            4.1.2. Add $c$ to $C$.
      4.2. If $C$ is not empty, return $C$.
```

Figure 1: Pseudo-code of the procedure.

Figure 1 shows the pseudocode of the algorithm used for inferring some semantic characteristics of unknown words. The main idea is the following: initially, we need a dictionary where words have to be classified according to some

[*] This work has been sponsored by CICYT, project numbers TIC2002-01948 and TIC2001-0685-C02-01.

‘Twas brillig, and the slithy toves Did gyre and gimble in the wabe:
All mimsy were the borogoves, And the mome raths outgrabe.

He took his vorpal sword in hand: Long time the manxome foe he sought –
So rested he by the Tumtum tree, And stood awhile in thought.

And, as in uffish thought he stood, The Jabberwock, with eyes of flame,
Came whiffling through the tulgey wood, And burbled as it came!

One, two! One, two! And through and through The vorpal blade went snicker-snack!
He left it dead, and with its head He went galumphing back.

”And, has thou slain the Jabberwock? Come to my arms, my beamish boy!
O frabjous day! Callooh! Callay!’ He chortled in his joy.

‘Twas brillig, and the slithy toves Did gyre and gimble in the wabe;
All mimsy were the borogoves, And the mome raths outgrabe.

Figure 2: Lewis Carroll’s Jabberwocky poem. The invented verbs are shown underlined.

semantic criterion. We also need the phonetic transcription of all the words from our dictionary. The length limit, set to three syllables per word, is motivated by the fact that, to our knowledge, most experiments of phonosemantics have been only tested with short words (typically one or two-syllable words) [Magnus, 2001; Marcos-Marín, 2001], and we doubt that it will be applicable to longer, usually more specific terms.

Given a new word, w, we automatically find the longest pattern of phonemes from w for which there are words in the dictionary following the same pattern. Among those, we keep the semantic category for which there are more words with the pattern (if any). Those words will be output as the result of the algorithm. Once a pattern with a given length has produced results, the search stops (step 4.2).

2.1 Experiment

An experiment has been performed for automatic acquisition of unknown verbs meaning, using Levin’s verb classes [Levin, 1993] as the semantic dictionary, and the invented verbs from Lewis Carroll’s poem *Jabberwocky* as test set (see Figure 2). The aim of the experiment is to check whether the meaning conveyed by the terms returned automatically for each of the invented verbs is consistent with the context of the poem. The text-to-phoneme transcription program used is t2p [Lenzo, 1997].

The first invented verb in the poem is *to gyre*, from which the phonetic transcript obtained is `JH AY R _`. The algorithm would look first for the pattern of all four phonemes in the dictionary, but it does not exists, as expected. Therefore, it removes one phoneme (there are four choices) and looks for the three-phoneme patterns. In this case, there are two results:

- `JH AY R .*`, which returns just one word *gyrate*, which belongs to Levin class 49 (body-internal motion).
- `.* AY R _`, which returns a total of eight verbs, all of which belong to different classes in Levin’s classification, so there is not a class that outstands.

As there are results for three-phoneme patterns, the search stops at that point, and the returned output is *gyrate*.

In a similar way, we obtain the classification of verb *to gimble*. The complete pattern (with six phonemes) does not appear in the dictionary. With five phonemes, pattern

`G .* M B AX L`

matches, but there is no outstanding class (all the classes have just one verb which matches the pattern), and we still cannot infer the correct meaning. Next, with just four phonemes, pattern

`.* M B AX L`

matches 16 verbs, among which class 51.3.2 (run verbs) is the most popular. Some verbs in this class are *to gambol* (to play boisterously) and *to ramble* or *to rumble*.

The third verb under study is *to whiffle*. Here, the only longest pattern which returns matching verbs is

`W IH .* AX L _`

for which there is only one verb as output: *to wiggle* (move to and fro), from classes 40.3.2, 47.3 and 49.

Finally, the last invented verb from the poem is *to galumph*. In this case, there is no pattern with more than 4 phonemes which matches any word in the dictionary. There are five three-phoneme patterns which match this verb, but in four of them there is not a prominent class, so the only class which outstands for a phoneme pattern is 37.3 (verbs of manner of speaking), with verbs *to grumble* and *to grunt*.

In the end, using the outputs mentioned above, a possible meaning for the poetry is that the “slithy toves where gyrating and either running or playing boisterously”; “the Jabberwock was running moving to and fro through the wood”; and “the protagonist went back grunting”.

3 Conclusions and future work

The results obtained show that, in this particular case, the verb meanings inferred are consistent with the contextual meaning of the text, which is very promising for the planned future investigation. The work described is preliminary, and can can be considered as the starting point inside a broader project which consists in ultimately testing empirically, with real data, whether the phonosemantic hypothesis really holds, and studying applications to assist in automatic ontology population for the semantic web. We envision the following lines for this work: (a) perform a thorough analysis of all of Levin’s verb classes, to identify the phonological patterns which are characteristic of some class of verbs; (b) extend the work to cover the other open-class words: adjectives, nouns and adverbs; and (c) apply this algorithm as an additional module to enrich current procedures for assisted ontology population.

References

[Genette, 1976] G. Genette. *Mimologiques*. Paris, Seuil, 1976. Also published as Mimiologics, Thas Morgan (trans.), University of Nebraska, Lincoln.

[Lenzo, 1997] K. Lenzo. t2p text-to-phoneme converter, 1997.

[Levin, 1993] B. Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. Univ. of Chicago Press, 1993.

[Magnus, 2001] M. Magnus. *What’s in a word? Studies in phonosemantics*. PhD thesis. Department of Linguistics, Science and Technology . NTNU Faculty of Arts, Norwegian University, 2001.

[Marcos-Marín, 2001] D. A. Marcos-Marín. Simbolismo en la estructura lingüística. In *Coloquio en las VI jornadas de lingüística*, 2001.

[Wallis, 1653] J. Wallis. *Grammatica linguae anglicanae*. Oxford, Hamburg, 1653.