Towards the Automatic Lemmatization of 16th Century Mexican Spanish: A Stemming Scheme for the CHEM

Alfonso Medina-Urrea

GIL, Instituto de Ingeniería, UNAM,
Ciudad Universitaria, 04510 Coyoacán, DF, Mexico
amedinau@ii.unam.mx

Abstract. Two of the problems that should arise when developing a stemming scheme for diachronic corpora are: (1) morphological systems of natural languages may vary throughout time, and these changes are normally not documented sufficiently; and (2) they exhibit very diverse orthographic characteristics. In this short paper, a stemming strategy for a diachronic corpus of Mexican Spanish is briefly described, which partially faces up to these problems. Success rates of the method are contrasted to those of a Porter stemmer.

1 Introduction

Diachronic corpora for the Spanish language have become available for various kinds of research. Two widely known corpora are the RAE's *Corpus diacrónico del español*, CORDE (http://www.rae.es/), and Mark Davies' *Corpus del español* (http://www.corpusdelespanol.org/). Recently, a first version of the *Corpus histórico del español de México*, CHEM (http://www.iling.unam.mx/chem/), became available to the public for the study of the Spanish used in Mexico from the arrival of Europeans to the 19th century.

Many tools for the exploitation and analysis of corpora require a lemmatization process, which is often reduced to simple stemming or graphical word truncation to eliminate inflections. Simple techniques such as the Porter algorithm [1] are regularly applied to corpora of many languages, but they require knowledge of their morphology. Fortunately, in comparison with other languages, Spanish morphology has changed relatively little during the last five centuries. So, a Porter stemmer for today's Spanish could presumably be applied to those centuries in order to accomplish inflection removal. However, given that techniques exist which can be used for stemming without having to code morphological knowledge into the algorithm, it is worthwhile to compare them to the Porter method in order to appreciate what scheme would be better for the CHEM.

In this short paper, the stemming strategy devised for this corpus is described and contrasted with an implementation of the Porter stemmer.[1] The strategy

[1] Various implementations of the Porter algorithm for Spanish are available (based on http://snowball.tartarus.org/). In this experiment a version for contemporary Spanish developed at GIL-IINGEN-UNAM, was used.

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 101–104, 2006.

proposed is based on automatic segmentation techniques previously tested for synchronic corpora of Spanish,[2] Czech, Chuj and Ralámuli.

2 Entropy and Economy Based Automatic Stemming

The stemmer basically examines each $n-gram$ of each graphical word, estimating uncertainty at each point, by measuring Shannon's entropy, and determining economy relations, by counting corpus evidence of syntagmatic and paradigmatic relations of word fragments. The techniques to accomplish this were sufficiently exposed for Czech and Chuj corpora in [2, 3, 4]. Also, they were presented with more detail for Spanish in [5]. In essence, each word segmentation yields, according to corpus evidence, entropy and economy measurements of how likely a morphological border is bound to occur at that segmentation. The highest values are expected to occur at the borders between bases and affixes, so they are taken as criteria to determine the best morphological segmentation within the graphical word. The stemmer simply eliminates the assumed suffix sequence.

3 Stemming a 16th Century Sample of Mexican Spanish

The target corpus for the stemming experiment is constituted by 95 CHEM documents from the 16th century. These documents comprise around 257,385 graphical token words, which correspond approximately to 15,834 graphical word types. Capitalized words were assumed to be proper names. These and words of length less than four were not stemmed. Upon examination of the documents, it becomes obvious that the orthographic idiosyncrasies of the 16th century cause referents to have multiple graphical forms (*e.g. admynistración, admynystracjon, administraçjon, adminystracjón*, etc.). Thus, although some unwanted homophony for short items may be introduced, it is clear that the text should be normalized in order to conflate, into a lesser number of graphical word types, several orthographical forms sharing a referent (*e.g. merced, merçed* → *merced; cantava, cantaba* → *cantaba; yndio, jndio* → *indio*, etc.). Therefore a simple set of rules to modify some characters was introduced to enhance grapheme-phoneme correspondence. These rules[3] appear in Table 1.

To be able to stem, the stemmer determines suffix sequences from the corpus estimating, as mentioned above, entropy and economy measurements. Using the corpus without the grapheme-phoneme correspondence rules, the method yielded 565 suffix strings. Then, given that stress marks distinguish verbal inflection morphemes (towards the end of words), the rules of Table 1 where applied to the corpus respecting last syllable stress marks (rendering 487 relevant suffix strings) and omitting all stress marks (rendering 470 suffix strings). The suffix

[2] An unpublished experiment indicates that these techniques perform, for contemporary Mexican Spanish, slightly better than the Porter method.

[3] There are, of course, problems that arise when applying these rules, but there is no space here to discuss them.

Table 1. Character modifications (grapheme-phoneme correspondence)

rules	ph.	contexts
'h' → ϵ	-	all contexts.
'v' → 'b'	[b]	all contexts.
'ch' → 'č'	[č]	all contexts.
'rr' → 'r̄'	[r̄]	all contexts.
'ç', 'z', 'c' → 's'[a]	[s]	'çe', 'çi'; every 'z'; 'ce', 'ci'.
'c', 'qu' → 'k'	[k]	'ca', 'que', 'qui', 'co', 'cu'.
'g' → 'j'	[h]	'ge', 'gi'.
'gu' → 'g'	[γ]	'gue', 'gui'.
'y' → 'i'	[i]	end of syllable preceded by vowel ('ay', 'ey', 'oy', 'uy'); or word beginning, before consonant ('yn', 'yd', etc.).
'j' → 'i'	[i]	between consonants or consonant and vowel; or word beginning before consonant ('jn', 'jd', etc.).
'r' → 'r̄'	[r̄]	beginning of word; or preceded by syllable ending with 'n', 'l' or 's'.

[a] By the 16th century, the Spanish stridents system was collapsing and most probably Castilian [θ] never made it to America; see [6].

sequences discovered by applying the rules and keeping last syllable stress marks were used for the stemming experiment, which, as mentioned above, consisted of finding the best segmentation of each word and then eliminating its right side.

4 Evaluation

For this evaluation, one of the 16th century documents was picked randomly. Then, both stemmers were applied to it and a specialist judged separately whether the segmentations were morphologically appropriate or not. Table 2 shows success rates for both stemmers applied to the selected document (12,424 token words). The CHEM column shows success percentages for the stemmer developed for this corpus. Since all errors occurred only once (*i.e.* in types of frequency one), the rate for types and the one for tokens is the same. The last two columns exhibit results for the Porter stemmer. As one would expect, rates improved somehow when the orthographic normalization rules devised for the CHEM (see Table 1) were applied before the Porter stemmer. Still, rates for the entropy-economy stemmer are much closer to 1.0.

Table 2. Success Rates for 16^{th} Century

	CHEM	**Porter**	
		without rules	with rules
types	0.9932	0.9328	0.9597
tokens	0.9932	0.8926	0.9328

5 Closing Remarks and Future Work

It is interesting that a Porter stemmer for contemporary Spanish would get such good rates when applied to a document belonging to an earlier stage of Spanish. This corroborates the observation that the morphology of this language has changed relatively little in the last centuries. One might ask, how far back can a method developed for one stage of a language be applied to earlier stages of the same language? The answer is obviously language dependent. More innovative languages like French or English have gone through considerably more changes in lesser time, so very likely Porter stemming based on their current states would be less adequate than it appears to be for Spanish. Another question would be, since the Porter stemmer obtained a type success rate of 0.96, why not just use such method, instead of going through the overhead of calculating entropies and finding affix economical relations, especially when the latter is more expensive? At least for this experiment, such expensive method did show an improvement, reaching a success rate of 0.99. It is not clear whether that small improvement is due to the robustness of the entropy-economy stemmer or simply to morphological differences between two stages of Spanish. At any rate, it is more desirable to apply the best existing method than to develop a Porter stemmer specifically designed for the 16th century. There is, however, lots of room for improvement. The stemming scheme presented will be improved and applied to the 17th, 18th and 19th centuries of the CHEM. This would be a first step towards lemmatization of all the corpus documents, a step necessary for the development of future exploitation tools.

Acknowledgments

The work reported in this paper has been supported by a DGAPA UNAM grant for PAPIIT Project IN400905.

References

1. Porter, M.F.: "An Algorithm for Suffix Stripping". Program **14**(3) (1980) 130–137
2. Medina-Urrea, A., Hlaváčová, J.: "Automatic Recognition of Czech Derivational Prefixes". In: Proceedings of CICLing 2005. Volume 3406 of Lecture Notes in Computer Science. Springer, Berlin/Heidelberg/New York (2005) 189–197
3. Medina-Urrea, A., Buenrostro Díaz, E.C.: "Características cuantitativas de la flexión verbal del chuj". Estudios de Lingüística Aplicada **38** (2003) 15–31
4. Medina-Urrea, A., Alvarado García, M.: "Análisis cuantitativo y cualitativo de la derivación léxica en ralámuli". In: Primer Coloquio Leonardo Manrique, Mexico, Conaculta-INAH (2004)
5. Medina-Urrea, A.: "Automatic Discovery of Affixes by Means of a Corpus: A Catalog of Spanish Affixes". Journal of Quantitative Linguistics **7**(2) (2000) 97–114
6. Harris, J.: "Historical Excursus: Reflexes of the Medieval Stridents". In: Spanish Phonology. MIT Press, Cambridge, Mass. (1969) 189–206