Homograph Ambiguity Resolution in Front-End Design for Portuguese TTS Systems

Daniela Braga. [1], *Luís Coelho*[2], *Fernando Gil V. Resende Jr.*[3]

[1] Microsoft Language Development Center, MSFT, Portugal
[2] Polytechnic Institute of Oporto, Portugal
[3] Federal University of Rio de Janeiro, Brazil

i-dbraga@microsoft.com, luiscoelho@eseig.ipp.pt; gil@lps.ufrj.br

Abstract

In this paper, a module for homograph disambiguation in Portuguese Text-to-Speech (TTS) is proposed. This module works with a part-of-speech (POS) parser, used to disambiguate homographs that belong to different parts-of-speech, and a semantic analyzer, used to disambiguate homographs which belong to the same part-of-speech. The proposed algorithms are meant to solve a significant part of homograph ambiguity in European Portuguese (EP) (106 homograph pairs so far). This system is ready to be integrated in a Letter-to-Sound (LTS) converter. The algorithms were trained and tested with different corpora. The obtained experimental results gave rise to 97.8% of accuracy rate. This methodology is also valid for Brazilian Portuguese (BP), since 95 homographs pairs are exactly the same as in EP. A comparison with a probabilistic approach was also done and results were discussed.

Index Terms: Text-to-Speech, homograph, disambiguation, Part-of-Speech parser, semantic analysis

1. Introduction

Homograph ambiguity is a well-known problem of difficult solution in TTS conversion. In Portuguese it is responsible for 0,62% of error rate in our LTS conversion, which means that in a 1000 sentence corpus, containing 9090 words, there are 57 mispronounced homograph words, because the output transcription is not the default one. Although this error rate may seem to be not very significant, it can be rather shocking from the user point of view.This is a complex issue, since it depends on morphological and syntactic information most of times. For instance, in <almoço> [al*'mosu] (*noun,* '*lunch*')/ <almoço>[al*'mOsu] (*verb*, '*I have lunch*'), the tonic vowel quality change is related with the POS each of the words belongs to, which is noun and verb respectively (SAMPA for EP is used for phonetic transcription, with one extension to represent the velar lateral consonant: e.g. <sal> ['sal*] ('*salt*')). Sometimes, homograph disambiguation can only be done by using semantic information, when words belong to the same POS (e.g. <besta> ['beSt6] (*noun*, '*beast*')/ <besta> ['bESt6] *(noun, 'crossbow'*)). The main work on homograph disambiguation for TTS systems can be found in [1], in which the author establishes a typology of homograph pairs for English, describes several techniques traditionally used to resolve homograph ambiguity (N-gram taggers, Bayesian classifiers and decision trees) and proposes a hybrid algorithm combining the best of the three described techniques. Regarding homograph disambiguation in EP TTS systems, the work of Ribeiro et al. [2], [3] can be mentioned. Although homograph disambiguation is not the main topic of this work, it shows that morphosyntactic information is responsible for the TTS good performance. In [2], Ribeiro et al. compare POS parsers, each using a hybrid approach (probabilistic and linguistically-rule based) and a probabilistic approach. Results seem to show a better performance of the hybrid approach. A table is shown with a typology of morphosyntactic ambiguities that influence the LTS converter. However, as no case of ambiguity is followed by examples, it is not clear whether the ambiguity cases are caused by homonym pairs or homograph pairs. Updates made in [3] seem to confirm that the morphosyntactic ambiguity that was analyzed is homonymic-based, which has little impact in LTS conversion, though it is very important to prosodic generation, concerning focus and prosodic group boundaries. In Seara et al. [4], [5], a POS parser is presented to predict the vowel quality of noun and verb forms in Brazilian Portuguese (BP) TTS systems. Although this is a very interesting work regarding homograph disambiguation, especially because the vowel quality variation along the verb inflection is shown to be predicted, it does not include homograph pairs whose disambiguation is semantically determined. Both morphosyntactic and semantic analysis were considered in Ferrari's et al. [6] approach for homograph disambiguation in BP TTS systems. The Cognitive Grammar framework was proposed and a corpora-based analysis was pursued in order to find the neighboring expected constructional schemes. This approach was tested with only one example. Although it is a very interesting approach, each existing homograph pair still needs a similar study. In this paper, we present a module, which is applied both to EP and BP, and that solves a large range of homograph ambiguity, either morphosyntactically or semantically determined. This paper is structured as following: in section 2, the architecture of the homograph disambiguation system is described; in section 3, tests and results are shown and a comparison with a probabilistic technique is discussed; in section 4, our main conclusions are presented.

2. System Architecture

The homograph disambiguation module can be seen as a part of the POS parser and it is integrated in the LTS converter, one of the most relevant modules of TTS Front-End component. In this section, we will describe the system design and its components, the methodology used, the proposed homograph typology and the corresponding algorithms.

2.1. Methodology

The first milestone of our work was collecting the maximum number of homographs, either through literature [7], [8], [9] or through performance results of our LTS converter [10]. We obtained a library of 106 homographs for EP, from which 95 of them are valid for BP. This was a rather difficult task, since a complete list of homographs for Portuguese is not available or published. The homograph collection is still proceeding, because the good performance of our module depends on the presence of the given homograph on the library. The second milestone was organizing the homographs into types, according with their grammatical category and phonetic alternation. Each type was given a code number to which an algorithm was matched. The third milestone was building algorithms with a set of conditions for each homograph type. In this stage, several libraries were collected [8], [9], [11] as described in section 2.2., and a POS parser was started. Disambiguation rules were trained with three different corpora: CETEM-Público (corpus containing newspaper language) [12], COMPARA (Portuguese-English aligned corpus containing literature translations) [13] and EUROPARL-Opus (multilingual aligned corpus containing transcriptions of European Parliament debates) [14]. This diversity in corpora was important to help us find different contexts for each homograph pair. In other words, one of the two possible readings of each homograph is more likely to be found in a certain corpus type than in another. For instance, in the pair <gosto>[o](‘*taste*’)/<gosto>[O] (‘*I like*’), it is very difficult to find the second form (verb in the first person singular) in a journalistic corpus like CETEM-Público, where subjectivity is avoided. However, it can be commonly found in literature texts, such as in COMPARA. The system was then tested with a different corpus as described in section 3.

2.2. Libraries

The following libraries were gathered: 1) Homographs library, containing 106 homograph pairs grouped in 24 types. 2) Closed POS library, containing the Parts of Speech that have a fixed number of items (pronouns, prepositions, adverbs, conjunctions, contractions, articles, numbers, determiners, interjections). 3) Morphemes library, containing noun, verb, adjective and adverb suffixes, prefixes, and Latin and Greek affixes. 4) Lemmas library, containing Portuguese Jspell dictionary [15] with about 34000 words morphologically annotated. 5) Irregular verbs library, containing the inflexion forms of the main irregular Portuguese verbs. 6) ‘Preparatory subject’ expressions library, containing expressions with verb to be in the third person + adjective followed by *that-clauses*: <é importante fazer/que se faça> (‘*it is important to do*’). 7) Restrict lexical combinations library (RLCL), containing idioms, proverbs or fixed expressions of one or more words, like <pregar um susto> (‘*to scare*’) or <boca do lobo> (‘*in the belly of the beast*’). This library is used in semantic analysis. 8) Wordnets library, designed under the Wordnet concept framework [16], containing words that are semantically and cognitively related with the homograph words. This library is also used in semantic analysis. Both Restrict lexical combinations and Wordnets libraries are corpora-based, since a Wordnet Project for Portuguese is still not available [17]. To build these libraries, CETEM-Público, COMPARA and EUROPARL-Opus corpora were used.

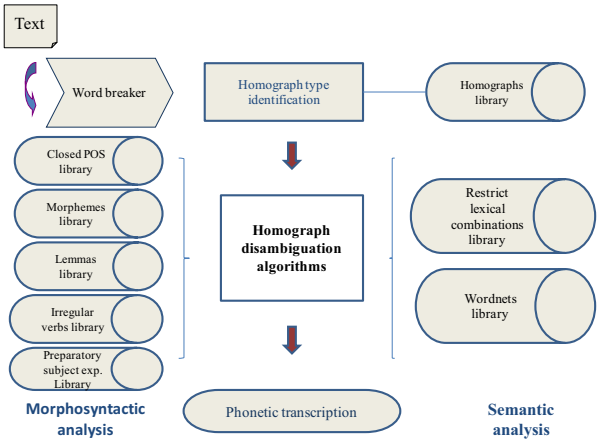

Figure 1: *Homograph disambiguation system architecture.*

In Figure 1, the system architecture is shown. The text is initially split into words. Then, the system performs a search for homographs. Once a given homograph is identified in the homographs library, it is conducted to its type and to the corresponding disambiguation algorithm. In Figure 1, libraries that are necessary to perform morphosyntactic analysis can be seen on the left, whereas libraries that are used for semantic analysis are placed on the right.

2.3. Homograph Typology

In Tables 1 and 2, the considered homograph typology is displayed. In Table 1, homographs belonging to different POS are shown, whereas homographs belonging to the same POS can be seen in Table 2. The homograph ambiguity of Table 1 is solved with morphosyntactic analysis, but semantic information is needed to resolve the ambiguity displayed in Table 2. The first criterion used to design our homograph typology was based on this distinction. The second criterion had to do with the POS opposition and the inherent vocalic alternation of each homograph pair. As the tables show, the most productive opposition occurs between Noun and Verb, from a morphological point of view, and between [e]/[E] and [o]/[O], from a phonetic point of view. Systematic evidence is that in Nouns, tonic vowels are typically closed, although in Verb forms tonic vowels are opened. Types 1 and 2 represent 68.3% of the total homographs library. Types 14, 15 and 20 need both morphosyntactic and semantic analysis, since they have three possible outputs. Types 12 and 24 establish a different vocalic alternation: it occurs in the pre-tonic syllable and, because of that, the opposition is [@]/[E].

Table 1. *Different POS homograph types.*

Type	POS opposition and vocalic alternation	Homographs
1	[e] Noun / [E] Verb	aceno, acerto, apelo, aperto, apreço, arrepelo, começo, concerto, conserto, desemprego, desespero, emprego, enredo, erro, esmero, espeto, flagelo, gelo, governo, interesse, interesses, modelo, peso, pego, rego, remo, selo, testo, zelo
2	[o] Noun/ [O] Verb	abono, aborto, acordo, adorno, aforro, almoço, arrojo, arroto, choro, conforto, consolo, consolo, contorno, controlo, coro, desgosto, despojo, destroço, encosto, endosso, esforço, estorvo,

		folgo, gosto, jogo, logro, namoro, olho, piloto, reforço, rogo, rolo, sopro, suborno, sufoco, toco, topo, torno, troco, troço
3	[o] Noun/ [O] Verb	rola, rolha, soma
4	[e] Verb / [E] Noun	colher, meta
5	[e] Contraction / [E] Verb	desses, deste, destes
6	[o] Verb/ [O] Adverb	fora
7	[e] Adj., N/ [E] Verb	seco, seca, secas
8	[o] Adj., N/ [O] Verb	boto
9	[e] Dem. / [E] Adj., N	este
10	[e] Verb / [E] Adj., N	leste
11	[o] Prep./ [O] Verb	sobre
12	[@] Verb./ [E] Noun	pegada
13	[o]Adj./[O]Noun	rota, rotas, tola, tolas
14	[o] Noun/ [O] Noun/ [O] Verb	corte, forma, formas, molho, soco
15	[e] Prep./ [e] Noun/ [E]Verb	cerca
16	[e] Noun/ [E] Verb, Noun	pega

Table 2. *Same POS homograph types.*

Type	POS opposition and vocalic alternation	Homographs
17	[e] Noun/ [E] Noun	besta
18	[e] Noun/ [E] Noun	sede
19	[e] Noun/ [E] Noun	medo
20	[e] Noun/ [E] Noun, Verb	termos
21	[o] Noun/ [O] Noun	cor
22	[o] Noun/ [O] Noun	lobo, lobos
23	[o] Noun/ [O] Noun	bola
24	[@] Noun/ [E] Noun	pregar

2.4. Disambiguation Algorithms

After the sentence is split into words, the system searches homograph candidates and matches them with its homograph library. If the system identifies a word as being a homograph, it matches it with a certain type. Each type has a corresponding algorithm with questions about the syntactic context of a given homograph. It uses the mentioned libraries to perform this task. In Figure 2, a decision algorithm is displayed as an example of Table 1's types of homographs. The symbols used are: P (current word), P-1 (last word), P-2 (second last word), P+1 (next word), DEM, IND, INT, POSS (demonstrative, indefinite, interrogative and possessive pronoun and determiner), ART_IND (indefinite article), CONT (contraction), PREP (preposition), P_REL (relative pronoun), P_PES_S/O (personal pronoun subject/object), CONJ_S (subordinate conjunction). The first set of questions leads to the most probable output according with the results obtained by corpora analysis. If the answer is negative, the system asks another set of questions and outputs the statistically less probable occurrence. If the answer is still negative, the system leads to the first output. In algorithms 17-24, the questions which are made are semantically-based. After being identified as "Type 24", for instance, the first set of questions is "is pr[E]gar_Wordnet present in current sentence, last sentence or next sentence?" or "does <pregar> belong to the pr[E]gar RLCL?". If the answer is yes, the output is ['prEgar]. If the answer is no, it leads to the second set of questions, similar to the first one but replacing the previous libraries by pr[e]gar_Wordnet and pr[e]gar RLCL. Again, if the answer is yes, the output is ['pregar], otherwise it is ['prEgar]. Every set of questions for each algorithm was manually programmed by a linguist who listed the possible syntactic contexts for each type of homograph pairs. During the algorithms' implementation, it was possible to reduce the number of types to 17, because the set of questions is mainly dependant on the POS and not on the phonetic output. More details on the questions and features involved in each algorithm can be seen in [18].

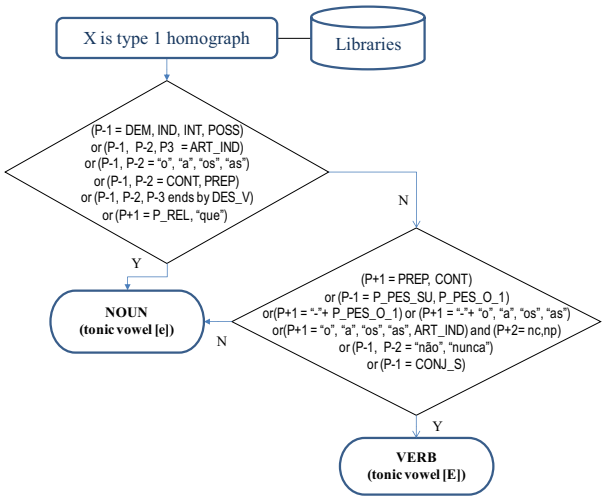

Figure 2: *T1 homograph disambiguation algorithm.*

3. Evaluation and Discussion

In this section, evaluation of the system is presented and a comparison with a different approach is carried out. Obtained results are discussed.

3.1. System Tests and Results

Table 3. *System test results.*

type	tested homograph	# occurrences	# errors	% errors
1	'erro'	59	0	0,0
2	'gosto'	67	5	7,4
3	'rola	3	0	0,0
4	'colher'	3	0	0,0
5	'desses'	64	0	0,0
6	'fora'	first 100 (in 252)	2	2,0
7	'seco'	4	0	0,0
8	'boto'	0	-	-
9	'este'	first 100 (in 1946)	0	0,0
10	'leste'	39	0	0,0
11	'sobre'	first 100 (in 2458)	0	0,0
12	'pegada'	0	-	-
13	'rota'	17	1	5,8
14	'forma'	first 100 (in 1154)	0	0,0
15	'cerca'	first 100 (in 1327)	11	11,0
16	'pega'	2	0	0,0
17	'besta'	0	-	-
18	'sede'	first 100 (in 398)	8	8,0
19	'medo'	92	0	0,0
20	'termos'	first 100 (in 523)	0	0,0
21	'cor'	34	0	0,0
22	'lobo'	1	0	0,0
23	'bola'	45	0	0,0
24	'pregar'	6	0	0,0
Total	-	1136	26	2,2

The system was tested with Natura-Diário do Minho corpus [19], containing 1738475 words of newspaper texts. We picked up a homograph of each type and ran the search system in the corpus. The obtained sentences were then

introduced in our LTS converter in order to check the homograph phonetic output. In Table 3, results of this procedure are shown. The global error rate is 2.2%. Much of the success rate has to do with the fact that if all the answers fail, the system outputs the first transcription. This result is very encouraging, considering that the checked homographs represent only 0.48% of the Natura-Diário do Minho corpus, which means that homographs have very little expression in this corpus. The nature of this corpus explains why some homographs cannot be found there. We could also confirm through the tests that the homograph outputs can be predicted by knowing the corpus nature. In newspaper corpora, for instance, it is more probable to have the noun output than any verbal inflection of 1st or 2nd persons. That is why three different corpora were needed for training.

3.2. Comparison with a Probabilistic Approach

In a previous unpublished work, as a first approach to the disambiguation problem, we have developed a stochastic HMM based system. Considering the traditional output-independence and Markov's assumption, for each word sequence (w) the system tries to maximize the probability of a tag sequence (t) as:

$$\hat{t}^k = \arg\max_{t^k} \prod_i^k P(w_i \mid t_i) P(t_i \mid t_{i-1}) \qquad (1)$$

Since it was not our objective to build a full POS tagger and our database was not completely annotated we used a simplified set of tags. Each word was manually labeled, by a linguist, as belonging to a closed class (according to the libraries described in 2.2), or as belonging to one of the interest morphological classes (those needed for disambiguation purposes as seen on Table 1) or as belonging to the "other" class, in a total of 25 different tags. A subset of corpus Natura-Diário do Minho was built with all the sentences containing homographs. The sub-corpus has been divided in two parts, one for training, with 80% of the occurrences for each homograph (total of 6751), and the remaining for testing (total of 1688). Counts of tag pair sequences were made for calculating the prior probabilities:

$$P(t_i \mid t_{i-1}) = C(t_{i-1} \mid t_i) / C(t_{i-1}) \qquad (2)$$

Good-Turing smoothing and Katz back-off have been used. Distinct HMMs have been trained for each homograph considering a four word neighborhood context. Using this simple system and considering the 1688 homograph occurrences reserved for testing we achieved an overall performance of 92.7%. This result is not far from the results published in the state of the art in POS tagging [2], [3] but in our case we have an easier task because the disambiguation output possibilities are seriously constrained. Some developments of the system are foreseen.

4. Conclusions

A linguistically-based module for homograph disambiguation applied to Portuguese TTS conversion was presented. This technique covers 106 homograph pairs, using 24 disambiguation algorithms. The system was tested and we obtained 97.8% of accuracy rate. Comparison with an HMM-based approach showed a better performance of our system using the same corpora for tests, although the used training corpora were different. The developed system proved to have 100% of accuracy rate in 16 types of homographs. We plan to conduct tests with other corpora. The application of this system to BP is ongoing. There is also an ongoing study on its application to Galician TTS conversion. This approach can be extended to other languages beyond Portuguese.

5. References

[1] Yarowsky, David, "Homograph disambiguation in Text-to-Speech Synthesis", Progress in Speech Synthesis (Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors), Springer, New York, 1996, pp. 159-174.

[2] Ribeiro, R. Oliveira, L.C., Trancoso, I., 2002, "Morphossyntactic Disambiguation for TTS Systems", Proc. of the 3rd Intl. Conf. on Language Resources and Evaluation, Volume V. pp. 1427-1431.

[3] Ribeiro, R. Oliveira, L.C., Trancoso, I., 2003, "Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese", PROPOR'200 3-6th Workshop on Computational Processing of the Portuguese Language, Springer-Verlag, Heidelberg, pp. 143-150.

[4] Seara, I., Kafka, S., Klein, S., Seara, R., 2001, "Considerações sobre os problemas de alternância vocálica das formas verbais do Português falado no Brasil para aplicação em um sistema de conversão Texto-Fala", SBrT 2001 – XIX. Simpósio Brasileiro de Telecomunicações, Fortaleza, Brazil.

[5] Seara, I., Kafka, S., Klein, S., Seara, R., 2002, "Alternância vocálica das formas verbais e nominais do Português Brasileiro para aplicação em conversão Texto-Fala", Revista da Sociedade Brasileira de Telecomunicações, vol. 17, nº 1, pp. 79-85.

[6] Ferrari, L., Barbosa, F., Resende Jr., F. G. V., 2003. "Construções gramaticais e sistemas de conversão texto-fala: o caso dos homógrafos", Proc. of the International Conference on Cognitive Linguistics. Braga, Portugal.

[7] Bergström, M., Reis, N. 1997, Prontuário ortográfico e guia da língua portuguesa. Lisboa, Editorial Notícias.

[8] Cunha, C., Cintra, L., 1992, Nova gramática do português contemporâneo, Lisboa, Sá da Costa.

[9] Estrela, E., Soares, M. A., Leitão, M. J, 2004, Saber escrever. Saber falar. Um guia completo para usar correctamente a língua portuguesa, Lisboa, D. Quixote.

[10] Braga, D., Coelho, L., Resende Jr., F. G. V., 2006, "A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese", VI International Telecommunications Symposium (ITS2006), Fortaleza-CE, Brazil.

[11] Nogueira, R. Sá, 1994, Dicionário de Verbos Portugueses Conjugados, Lisboa, Clássica Editora.

[12] Cetem_Público, http://acdc.linguateca.pt/cetempublico/

[13] COMPARA, http://www.linguateca.pt/COMPARA/

[14] EUROPARL, http://logos.uio.no/cgi-bin/opus/opuscqp.pl?corpus=EUROPARL;lang=pt

[15] Jspell, http://natura.di.uminho.pt/wiki/index.cgi?jspell.

[16] Wordnet, http://wordnet.princeton.edu/.

[17] PT Wordnet, http://www.clul.ul.pt/clg/projectos/WordNet.PT-I.html.

[18] Braga, D., Marques, M.A. 2007. "Desambiguador de homógrafos heterófonos para sistemas de conversão Texto-Fala em Português", in *Revista Diacrítica – Série Ciências da Linguagem*, Braga, Univ. Minho, Portugal.

[19] Natura-Diário do Minho: http://www.linguateca.pt/.