Parts of Speech Recognition System for the Text-Based Polish Speech Synthesizer

B. Piórkowska, J. Rafałko, Ł. Kalinowski, K. Pąk

Institute of Computer Sciences, University of Bialystok,
Sosnowa str. 64, 15-887 Bialystok, Poland.
boncia@wp.pl

Abstract

The article presents two ways of automatic recognition what part of speech a given word is. It is crucial to determine which words belong to subject and which to predicate. This information will greatly enhance proper sentence intonation because, as observations have proved, the most frequently stressed words in a sentence are the ones from the subject group. The two recognition algorithms, their good points and drawbacks, are discussed in the article. Apart from that, there is also an evaluation of how the program based on these algorithms works.

1. Introduction

Proper synthesized sentences intonation is crucial for the speech to sound natural and not be monotonous. The conducted research proved that subject was most frequently stressed. In order for the program to determine which words are from the subject group it has to know what part of speech each word in this sentence is. It would not be a good solution to create a database of all words in Polish (over 4 million words) with information what part of speech each word is; especially in a situation where time of data gathering is crucial. That is why proper algorithm elaboration was needed; an algorithm which quick and precisely would match a word with a part of speech. Two algorithms were created and tested. One mostly based on inflectional characteristics, the other one uses the word written in its phonemic form.

2. Parts of Speech in Polish

In Polish words are classified into groups. In addition, some of them appear in different conjugation. And so:

- nouns (e.g. *kot, Kraków* – cat, Cracow) – inflect for number (singular, plural) and case (nominative, genitive, dative, accusative, instrumental, locative, vocative);
- verbs (e.g. *pisze, bawi* – writes, plays) – inflect for number, tense (past, present, future), mood (indicative, conditional, imperative), aspect (perfective, imperfective, biaspectual), voice (active, passive, reflexive), person (first, second, third) and gender (singular masculine, feminine, neuter, plural masculine personal, non-masculine);
- adjectives (e.g. *dobry, zdrów* – good, well) – inflect for number and gender;
- adverbs (e.g. *dobrze, najładniej* – well, prettiest) – do not inflect;
- numerals (e.g. *jeden, setny* – one, hundredth) – some inflect for case and gender;
- pronouns – not all inflect. Inflection is the same as the part of speech it substitutes;
- prepositions (e.g. *przy stole, za stołem* – at the table, behind the table) – do not inflect;
- conjunctions (e.g. *i, lub* – and, or) – do not inflect;
- particles (e.g. *chyba, nie* – I guess, no) – do not inflect;
- exclamations (e.g. *ach, ej* – ah!, hey!) – do not inflect;

Parts of speech are divided into conjugation groups. There are eighteen such groups for verbs, four for adjectives and fifteen for nouns. Depending on conjugation type words have various inflectional endings. In order to state what part of speech a word is, first of all the stem had to be separated from the inflectional ending because basing on this form, identification of the part of speech is possible. However, for some words endings are the same as for other parts of speech. Some of these endings are:

- ~a (nouns – *krata, moda* (check, fashion), verbs – *czyta, zbada* (reads, will examine)), ~nia (nouns – *kania, jedzenia* – (parasol mushroom, food), verbs – *oddzwania, kłania* (calls back, bows)) etc.;
- ~i (nouns – *rogi, drwali* (horns, woodcutters), adjectives – *nagi, tani, ośli* (naked, cheap, donkey's), verbs – *ocygani, chwali* (will diddle, praises), ~y (nouns – *tory, maliny* (railway, raspberries), adjectives – *siny, dumny* (strong, proud)) etc.

Moreover, there are certain words which are different parts of speech depending on the context. For example:

- the word ***myśli*** (thinks, thoughts) in the sentence: *On **myśli**, że jest siódma rano* (He thinks it is seven o'clock in the morning) is a verb, whereas in the sentence: *Po głowie biegały mi różne **myśli*** (I had many thoughts in my head) it is a noun;
- the word ***drogi*** (expensive, way) in the sentence: *To jest **drogi** zegarek* (This is an expensive watch) is an adjective, whereas in the sentence: *Nie mogłem znaleźć właściwej **drogi*** (I could not find the right way) it is a noun.

3. Recognition Mechanism

There are two ways to find out what part of speech a word is: using words database or using a program (one based on an algorithm). However, in order to compare the effectiveness of both ways, a database with information concerning what part of speech each word is had to be created (a database of about 4 million words, the word occurrent in every existing for this word of conjugation). Although creating the database was not a problem, matching each word with the relevant part of speech would be an immensely time-consuming duty.

In order to recognize words like numerals, conjunctions, pronouns, prepositions, particles and exclamations, their database can be successfully used; it is due to the fact there are not many of them. That is why the problem is narrowed to identifying nouns, verbs, adjectives and adverbs. In case of verbs which are participles it has to be stated what kind of a participle it is. Moreover, verbs which are verbal nouns (e.g. *słuchanie* – listening) need to be determined. This additional division is related with the role the verb might have in a sentence.

3.1. Phonemic Transcription-Based Algorithm

The first one of them is very easy and fast. Having access to the phonemic transcription of words with marked *stress (the + symbol appears after the stressed vowel) e.g. celować->celo+wać* (to aim), *obciążony->općąże+ńi* (loaded), we can divide the word into "*before_the_stress_part+after_the_stress_part"*.
Distinguishing the ending of a word, which will allow to identify the part of speech, comes down to finding the "*+after_the_stress_part"*. In order to work properly the algorithm requires the phonemic transcription of words (this is not a problem due to the fact that such transcription is generated during speech synthesizing) and the database with endings of particular parts of speech. Such database stores around 450 different forms. Firstly, the ending of a word is taken (*after_the_stress_part*) and after that it is compared to the model from the database. Because of the number of occurrence it first checks if the word is a verb, then an adjective and then an adverb. If the examined word does not qualify to any of these groups, it is classified as a noun.

3.2. Polish Grammar-Based Algorithm

The second algorithm is based on the notion that in Polish verbs have the most inflections and their inflectional endings are the longest as compared to other parts of speech; another notion is that if a given word appears in the database, various inflections are provided.

A verb has eighteen basic conjugation forms, each of which has, on average, fifty inflection forms of a verb in a given form (maximum endings per one inflection type is seventy-three). The first part of words' interpretation is checking for each word the existence of a string of letters at its end corresponding to the proper inflectional ending of a verb (meaning one out of around 1100 in the database). It is clearly seen that if each word, in which the verb ending was found, was labelled as a verb, around 70% of all words would be verbs. In order to remove this fault from the algorithm, each word was assigned the maximum length of an ending. This way helps to find the stem of a word. However, the stem determined this way does not necessarily have to be the real one.

- **Verb Recognition**

The main part of verb recognition algorithm is based on this observation: **due to the fact that the database has almost all forms of inflections, it is possible to verify the hypothesis, whether the recognized stem is indeed the stem, by verifying the amount of designating this stem**. So, for a given word, let us name it X, all the words are searched (in the database which is alphabetically sorted) which meet the following conditions:

- their beginning matches the stem
- their length is the same
- the recognized stem length is in accordance with the length in X

Out of the designated group of words each of them and every of the eighteen forms of verb's inflections is examined as to if the word was recognized as the word with the ending from the list of eighteen inflections. Then, for every group of inflection, the number of words found in a given group and the form which has the most words, are being designated. It will be marked F.

However, it turned out that this hypothesis leads to some errors (5% margin). There are adjectives and adverbs which are inflected like some participles and with stem designated this way, nouns which are created from adjectives and adverbs were recognized as the infinitive form of verbs.

In order to erase the error, a rule has been added that amongst the recognized endings of F, at least five must appear which are related with the verbs which are not participles or verbal nouns. Such hypothesis turned to be extraordinarily accurate.

- **Adjective Recognition**

The beginning of this process is similar to the abovementioned one, but this time, twelve basic adjective forms with corresponding endings are used. In a similar way probable stems are determined (of course, words which are not recognized as verbs in the first phase are verified). Here, however, examining only the amount of endings of a fixed length proved to be sufficient. This amount depends on the adjective's inflection form. Moreover, participles found in the first phase were used – if the found word has the same stem as the previously found participle, then, regardless of the amount of particular endings, an assumption is made that the word has been recognized as an adjective.

- **Noun Recognition**

Nouns are recognized in a similar way to adjectives. Fourteen inflectional forms have been singled out and the minimum occurrence number was empirically set to be at least three. Furthermore, words whose stem is the same as of the previously recognized verbal nouns, are treated as nouns.

4. Algorithms Tests Results

The algorithms were tested on a database of 8295 words. They (the algorithms) were analysed paying special attention to correctness of recognized words. The first algorithm recognized all the words and did it fast, but its correctness was 87%. It reflects the fact that some words have identical inflectional endings for different parts of speech. In other words, certain group of *after_the_stress_part* endings is identical for various parts of speech.

The second algorithm was better; 98% of recognized words were identified correctly. However, not all words were recognized. 105 (1.3%) out the total 8295 words were left unclassified. The unrecognized words were mostly the ones inaccurately classified by the first algorithm. If we treat these words as incorrectly classified, the algorithm's correctness is not less than 96%. It is a very good result. The drawback is its speed. It works three-and-a-half times slower than the first one.

5. Conclusions

In order to locate the subject of a sentence and to know how the sentence is built, one needs to recognize what parts of speech the sentence is consisted of. This information will allow to determine where sentence stress should be. Elaborating these two algorithms will prove which method is better. Will it be the program recognition of parts of speech which makes many mistakes but is fast and does not require the database of all words, or will it be identification through finding the word in the right database, which is far more time-consuming, yet almost flawless? One might think that program identification will be better because it is faster. However, the mistake it can produce may be so serious that subject in a sentence will be determined incorrectly. The result could be wrong intonation of the synthesized sentences and this should be avoided at all costs

6. Acknowledgements

This paper was supported by the EUROPEAN COMMISSIN under grant INTAS Ref. number 04-77-7404. The author wish to express their thanks for the support.

7. References

[1] Shpilewski E., Piorkowska B, Rafalko J., Lobanov B, *Kiselov* V., Tsirulnik L. "Polish TTS in Multi-Voice Slavonic Languages Speech Synthesis System", Proceedings of the 9th International Conference "Speech and Computer" – SPECOM'2004, Saint Petersburg, 2004.

[2] Piorkowska B., Rafalko J., Shpilewski E. „Conversion of Textual Information to Speech for Polish Language", Proceedings of the 4th International Conference on Computer Recogniotion Systems – CORES'2005, Wroclaw, 2005.

[3] Piorkowska B., Rafalko J., Lesinski W., Shpilewski E., „Sentence Intonation for Polish Language", Speech Analysis, Synthesis and Recognition. Applications of Phonetics, Krakow, 2005.