CHAPTER 3

PROPOSED METHOD

3.1 Overview

MRAS is based on a supervised learning approach that relies on knowledge acquired from a leveled corpora. In designing MRAS we followed the steps illustrated in Figure 3.1 and discussed below.

MRAS receives two different inputs: a collection C of documents for each of which a readability label is already assigned, and a document d which its readability is unknown for MRAS and, thus, will be predicted. Both inputs are taken through a preprocessing step described in section 3.3, which cleans, filters and normalizes their content, to prepare them for the feature extraction step described in section 3.4. These features, serve as a numeric representation for each document. MRAS is capable of learning patterns over the representations extracted from each document in C and use these patters to predict readability scores for new unlabelled documents such as d.

3.2 Tools

Whether for preprocessing or for feature extraction, MRAS takes advantage of several existing tools and techniques. Each of those tools is described below.

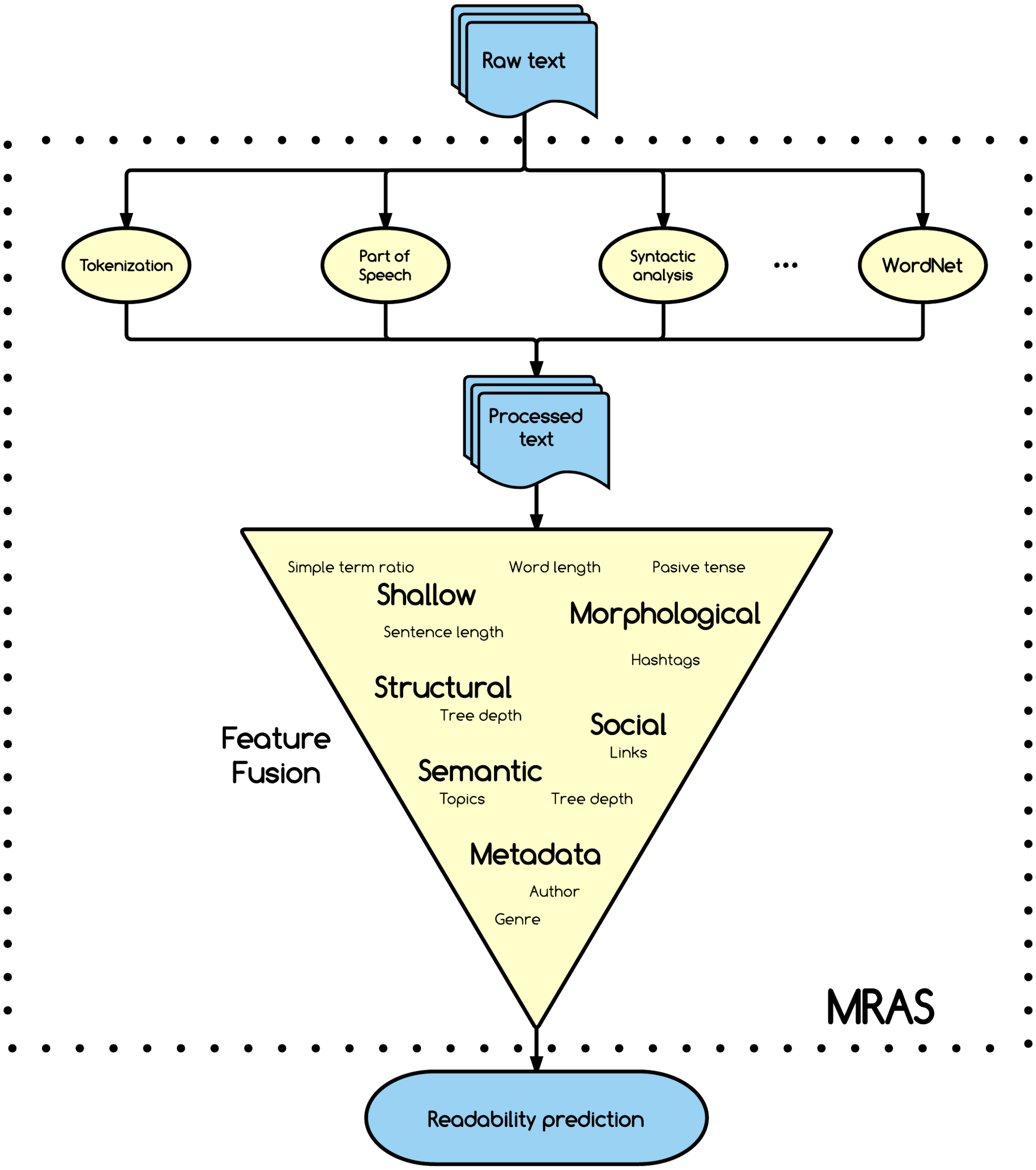

Figure 3.1: MRAS

Freeling NLP

Freeling [43] is a NLP tookit developed for the easing various natural language analysis tasks. Freeling includes, but is not limited to, tokenization, PoS tagging, syntactic parsing, dependency parsing and semantic labelling. Furthermore, Freeling is, to the best of our knowledge, the only tool kit supporting 14 languages with this depth of analysis, supporting Asturian (as), Catalan (ca), German (de), English (en), French (fr), Galician (gl), Croatian (hr), Italian (it), Norwegian (nb), Portuguese (pt), Russian (ru), Slovene (sl), Spanish (es), and Welsh (cy).

SyntaxNet

should I mention why not this one?

Katea

Katea is set of NLP tools developed for Basque. Katea is composed by Morpheus [11] (morpho-syntactic analysis), eustagger [15] (lemmatization and syntactic function indentification), eihera [14] (named entity detection), ixati [11] (shallow parsing) and maltixa [18] (dependency parsing).

WordNet

Wordnet [38] is a lexical database for English where terms, i.e., nouns, verbs and adjectives, are grouped into sets of synonyms, expressing different concepts. These concepts are related by each other by several semantic relationships, such as hyperonymie or hyponimie. An example of this structure can be seen in figure 3.2, where *motor vehicle* has *vehicle* as hypernym and *car* and *truck* as hyponyms. Note that this structure forms a tree. This fact will be used later, for building some features.

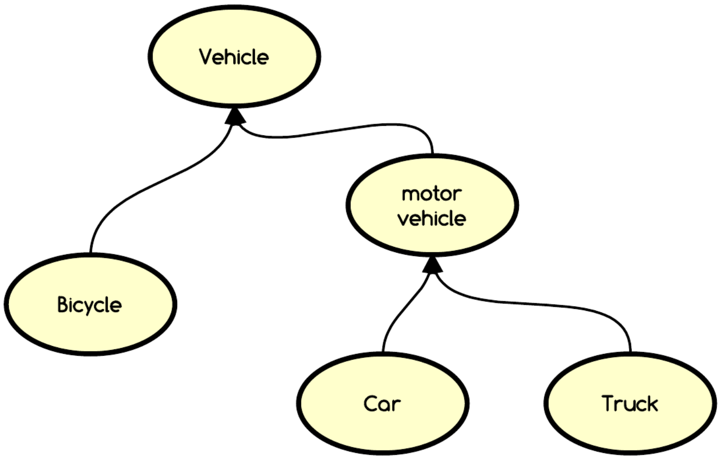

Figure 3.2: Wordnet example

3.3 Preprocessing

The preprocessing step is the one that takes place first for any document handled by MRAS. The aim of preprocessing is to identify the document to decide further process MRAS needs to perform over it, and to prepare the document for future feature extraction. More detail of mentioned processes is given below.

3.3.1 Document type detection

One of the main features of MRAS is its versatility, since MRAS is capable to predict readability values for documents of different format, length and language. Given this variety of documents, each of them cannot be treated the same way, different strategies need to be applied for different texts. Therefore, each document used by MRAS is classified using 3 criteria: format, length and language.

should we go deeper? explain how each criteria is determined?

3.3.2 Tokenization

Tokenization is the process of splitting a text into smaller parts, i.e. tokens. A token represents each sensical part of a text, which usually corresponds to a term, a number, or a punctuation mark. However, sometimes tokens can be formed by a combination of the previous, i.e., *aren't* or *people's*. An example of how a sentence is tokenized can be seen in Table 3.1.

<table>
<tr><td colspan="6">Did they win the olimpics?</td></tr>
<tr><td>did</td><td>they</td><td>win</td><td>the</td><td>olimpics</td><td>?</td></tr>
</table>

Table 3.1: Tokenization example

3.3.3 Part of Speech Tagging

Part of Speech Tagging is the process of labelling each token with a tag that represent the function each token has in a sentence. PoS tags usually differ from language to language[1], however, the most predominant tags, such as verb, adjective or noun, exist among all the languages. An example of Part of Speech tagging can be seen in Table 3.3.

did	they	win	the	olimpics	?
Verb	Pronoun	Verb	Determiner	Noun	Symbol

Table 3.2: Part of Speech tagging example

3.3.4 Shallow parsing

Shallow parsing, also called chunking, refers to the process of grouping tokens into chunks. A chunk consists usually a small phrase of about 1 to 4 terms. Those terms

[1] As an example, the Penn Treebank project defines 36 PoS tags for English, which can be seen here https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

are somehow connected to each other and together express a senseful concept. There are two types of chunks, depending if they express a noun or verb phrase. An example of a shallow parsing of a sentence can be seen in figure 3.3.

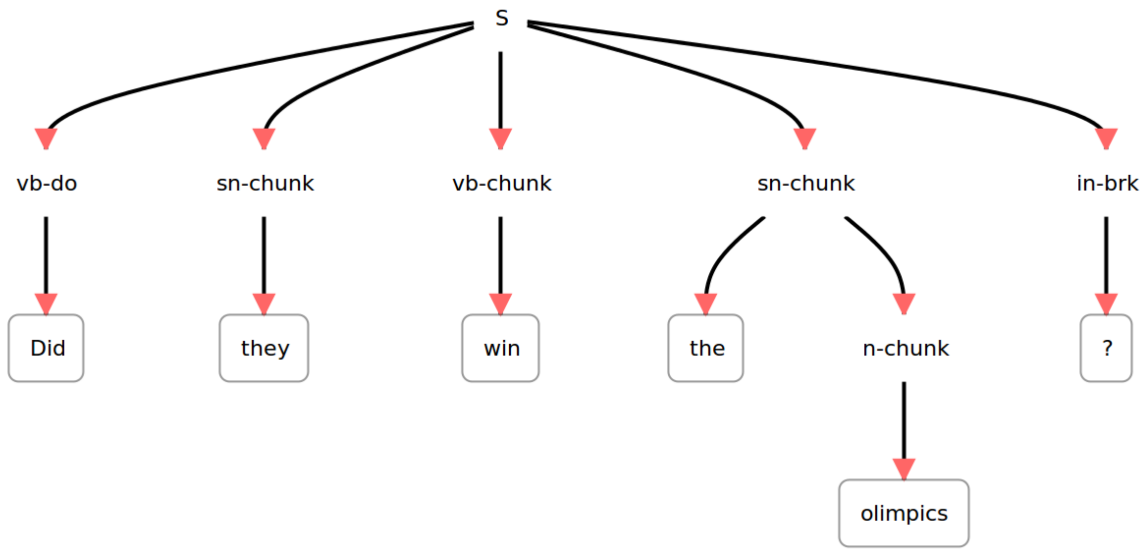

Figure 3.3: Shallow parsing example

3.3.5 Dependency parsing

Dependency parsing goes further than shallow parsing, determining relationships between tokens rather than just grouping them. Given these relationships, a dependency tree is generated, which usually has a root node representing the main verb of the sentence, which has the subject and objects of the sentence as children. An example of a dependency parsed sentence can be seen in figure 3.4.

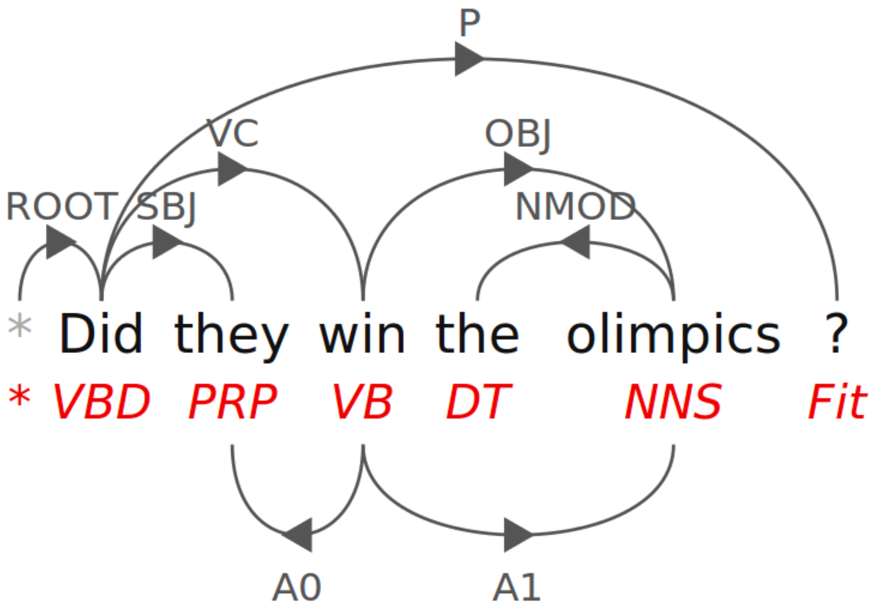

Figure 3.4: Dependency parsing example

3.3.6 Named entity detection

A named entity is a token or group of tokens that represent and known entity such as a person, a location, or an organization. Depending on the complexity of the tool that performs this analysis, those entities can also be linked to a knowledge base such as Dbpedia [36] where more structured information about the entity can be found.

Usain	Bolt	won	the	race	in	Rio	de	Janeiro	.
person	person					location	location	location	

Table 3.3: Named entity detection example