Identification of Direct/Indirect Discourse in Children's Stories

Nuno J. Mamede

L^2F – Spoken Language Systems Lab
INESC-ID Lisboa / IST, R. Alves Redol 9, 1000-029 Lisboa, Portugal
Nuno.Mamede@l2f.inesc-id.pt
http://www.l2f.inesc.pt

Abstract. The automatic identification of direct and indirect discourses is a topic not yet explored in Natural Language Processing. We developed the DID system that when applied to children stories identifies the discourses, relative to the narrator (indirect discourse) or to the characters taking part in the story (direct discourse). This automation can be advantageous, namely when it is necessary to tag the stories that should be handled by an automatic story teller [1].

1 Introduction

Children stories have some intrinsic magic that captures the attention of any reader. This magic is transmitted by intervenient characters and by the narrator that contributes to the comprehension and emphasis of the fables. Inherent to this theme emerges the direct and indirect discourse apprehension by the human reader that corresponds to character and narrator, respectively. This work deals with the separation between direct and indirect discourses of children fables. This separation leads to identification of the different agents responsible by the speech of expressed phrases under direct (and indirect) discourse form. This distinction is expressed in a final document with tags associated with each character. For example, starting with the following excerpt of a story (Although all examples being in english, our system only handles portuguese texts).

```
They arrived at the lake. The boy waved to them, smiling.
Come, it is really good!
```

We intend to identify the text that can be associated with each character of the story:

```
<person name="narrator">
   They arrived at the lake. The boy waved to them, smiling.
</person>
<person name="boy">
   Come, it is really good
</person>
```

2 Background

In order to apply DID it is necessary to resort, first, to two other systems: Smorph [4], a morphological analyzer, and PasMo [2], which divides the text into paragraphs and transforms word tags. Thus, the story texts are first submitted to Smorph and then PasMo, which produces XML documents, following this DTD:

N.J. Mamede et al. (Eds.): PROPOR 2003, LNAI 2721, pp. 175–178, 2003.

```
<!ELEMENT text (phrase)*>
<!ELEMENT phrase (hypothesis)*>
<!ATTLIST phrase num CDATA "1">
<!ELEMENT hypothesis (word)+ >
<!ATTLIST hypothesis num CDATA "1">
<!ELEMENT word (classification)* >
<!ATTLIST word name CDATA #REQUIRED>
<!ELEMENT classification (#PCDATA)>
<!ATTLIST classification root CDATA #REQUIRED>
<!ATTLIST classification c CDATA #REQUIRED>
```

3 Solution

First, we collected a set of children stories, all of them by Portuguese authors, and divided into a training set(eleven stories), and a test set (four stories). From the hand inspection of the texts, we extracted twelve heuristics:

Heuristic 1 A dash at the beginning of a paragraph identifies a direct discourse;
Heuristic 2 A paragraph mark after a colon suggests the paragraph corresponds to a character (direct discourse);
Heuristic 3 If a paragraph has a question mark in the end then probably it belongs to a character because the narrator uses less this type of mark. A character can be questioning someone else. However this heuristic depends on the type of paragraph;
Heuristic 4 The exclamation mark in the end of a paragraph identifies a direct discourse, with some probability. This heuristic follows the reasoning of H3;
Heuristic 5 The personal or possessive pronouns in the 1st or 2nd person indicate that we are in the presence of a direct discourse;
Heuristic 6 Verbs in past tense, present, future or imperfect tense are characteristics of direct discourse because they are verbs directed to characters;
Heuristic 7 The usage of inverted commas can indicate the speech of a character, but generally it is the narrator imitating the character and not the character speaking about himself/herself;
Heuristic 8 The usage of tense adverbs (tomorrow, today, yesterday, etc.) can identify a direct discourse;
Heuristic 9 If next to a direct discourse there is a dash, then a little text and another dash, then the next text excerpt probably belongs to a character;
Heuristic 10 The imperfect tense verbs that can be expressed in the same way for a character and for a narrator just lead to a direct discourse when there is a personal pronoun corresponding to a character;
Heuristic 11 In the phrase, if there is a text excerpt between two dashes where a declarative verb exists (declare, say, ask, etc.) in the third person, then we can say that a character expresses the text excerpt appearing before the left dash;
Heuristic 12 The use of interjections identifies a direct discourse because only characters use them.

However, when DID was implemented we needed to operate some changes to these heuristics, namely:

- Heuristic 3 and Heuristic 4 have different trust values when some question or exclamation mark appears in the middle of a paragraph or in the end. When in the middle, the trust value must be lower, and when at the end, the trust value must be higher. So, these heuristics have two trust values (a minimum and a maximum).
- Heuristic 5 and 6 have been combined, because DID's input has many ambiguities.
- Heuristic 7 is a neutral heuristic so it is not applied to direct discourse.

The input to DID is PasMo's output. DID analyses the text paragraph by paragraph. Heuristics are then applied to each one. After processing the whole text, DID returns an XML document, in VHML format [3], that contains all the identified discourses accordingly to the tags supported by this language.

DID followed the Theory of Confirmation to get the degree of trust with which identified direct discourse: the user can define the trust to associate with each heuristic and also the value of its threshold, which defines the limit between success and failure. Thus, we can say that DID works like an expert system.

Table 1. Results of DID measured by DID-Verify

Story	Correct	Incorrect	Success rate
O Gato das Botas	28	0	100%
O Macaco do Rabo Cortado	48	0	100%
O Capuchinho Vermelho	41	1	97%
Os Trés Porquinhos	28	1	96%
Lisboa 2050	147	6	96%
A Branca de Neve	43	2	95%
Ideias do Can‡rio	41	2	95%
Anita no Hospital	102	11	90%
Os Cinco e as Passagens Secretas	131	19	87%
A Bela e o Monstro	31	6	83%
O Bando dos Quatro: A Torre Maldita (Cap'tulo 1)	70	40	63%

4 Discussion

In order to check the capabilities of DID system, we developed a new system: DID-Verify, which is responsible for the comparison between DID's output and the idealized result. This comparison verifies whether discourses were well identified by DID and also shows the number of times that each heuristic was applied.

After analyzing the results obtained with the training set, we can easily infer that the best results are obtained for the children stories (e.g. O Gato das Botas, O Macaco do Rabo Cortado), what can be explained by the fact that characters are mainly identified by Heuristic 1. The worst result is obtained with the story ÒO Bando dos QuatroÓ, because here the narrator is also a character of the story, leading to an ambiguous agent: sometimes speaking like a narrator and others like a character. DID is not prepared to treat this ambiguity. Two children stories achieved 100% successful results, confirming the good performance of DID as a tagger for a Story Teller System under development by other researchers of our research institute. The result obtained for the story ÒLisboa 2050Ó must be heightened because this story has a large number of discourses and DID performs a 96% successful result! Summarizing the results, DID obtains an average of 89% of success showing that the results are similar to the projected objectives.

Analyzing the test set, all the results surpass 80% of success with an average of 92%. That is very reasonable for a set of texts that was not used to train the DID system. This result also shows that DID has a fine performance in different types of stories.

Examining the results obtained by DID-Verify with the test set, we obtained the 2, which shows the performance of each heuristic. Here we conclude that Heuristic 1 is the most applied, identifying a larger number of discourses correctly. Heuristic 5 and Heuristic 6 also lead to good results. Heuristic 2 never fails but was only applied six times. Heuristic 4 is the one that leads to more mistakes, because the exclamation mark is many times used in narration discourses. Generally, all the heuristics have a high success rate.

Table 2. Analysis of correctness

Heuristic	N Successes	N Failures
H1	188	2
H2	6	0
H3	59	1
H4	37	3
H5	81	2
H6	70	1
H8	7	1
H12	17	1

5 Future Work

To improve the DID system, we plan to (i) define associations of words and expressions to help identify some type of story characters; (ii) define a set of verbs that cannot be expressed by a narrator; and (iii) to use a morphossyntatic disambiguator to handle the ambiguous word classifications.

DID-Names is system that already started to be developed, and will be able to identify the character(s) that is(are) responsible for each direct discourse. Later on, we would like to identify the gesture and emotions as well as the environment where each scene takes place.

References

1. A. Silva, M. Vala, and A. Paiva, Papous: The Virtual Storyteller, Intelligent Virtual Agents, 3rd International Workshop, Madrid, Spain, 171–181, Springer-Verlag LNAI 2190 (2001).
2. J. Paulo, M. Correia, N. Mamede, C. Hagège: “Using Morphological, Syntactical and Statistical Information for Automatic Term Acquisition", in Proceedings of the PorTAL – Portugal for Natural Language Processing, Faro, Portugal, Springer-Verlag, 219-227 (2002).
3. C. Gustavson, L. Strindlund, W. Emma, S. Beard, H. Quoc, A. Marriot, J. Stallo, VHML Working Draft v0.2 (2001).
4. S. Ait-Mokhtar: L’analyse présyntaxique en une étape. Thèse de Doctorat, Université Blaise Pascal, GRIL, Clermont-Derrand (1998).