Direct and Indirect Discourse Separation (DID – Discourse IDentifier)

Diogo Barbosa, Ricardo Costa, Nuno J. Mamede

(L²F INESC-ID / IST)
Laboratório De Sistemas de Língua Falada
Rua Alves Redol 9, 1000-029 Lisboa, Portugal
{diogobarbosa,rflcosta}@hotmail.com, Nuno.Mamede@inesc-id.pt

Abstract – The automatic separation between direct and indirect discourses is a subject not yet explored in Natural Language Processing. We developed the DID system that can be applied to children stories: identifies the discourses relative to narrator (indirect discourse) or to characters (direct discourse). This automation can be advantageous, namely when it is necessary to tag the stories that should be presented by a story teller.

1 Problem posting

Children stories have some intrinsic magic that captivates the attention of any reader. This magic is transmitted by intervenient characters and by 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. Given the relevance of narrator and character roles, 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 to each character. For example, starting with the following excerpt of a story,

They arrived to the lake. The boy waved to them, smiling.
Come, its really good!

we pretend to identify the text that can be associated with each character of the story:

```
<person name=”narrator”>
   They arrived to the lake.
   The boy waved to them, smiling.
</person>
<person name =”boy”>
   Come, its really good
</person>
```

2 Background

In order to apply DID it's necessary to resort in first place to two other systems: Smorph and PASMO[2]. Smorph is a morphological analyzer that classifies each word in a set of morphological hypothesis. This classification is then used by PasMo (Pós-Análise Morfológica) which separates the text by paragraphs and transforms the word tags.

Thus, the story texts are first submitted to Smorph and then PasMo, which produces XML. Figure 1 contains the corresponding DTD.

```
<!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>
```

Figure. 1 – DID's input DTD

Title	Author	Number of characters (letters)	Number of words	Number of paragraphs
Anita no Hospital	Gilbert Delahaye	5469	1164	98
Os cinco e as passagens secretas	Enid Blyton	7737	1692	67
O bando dos quatro: A torre maldita (capítulo número 1)	João Aguiar	9290	2018	73
O macaco do rabo cortado		2575	578	46
O gato das botas		1885	413	27
Os três porquinhos		2035	492	27
A Branca de Neve		4384	1015	37
A Bela e o Monstro		2517	547	33
O Capuchinho Vermelho		2393	552	43
Lisboa 2050		15944	3464	128
Ideias do Canário	Machado de Assis	7993	1672	45
Pinóquio		2752	547	37
O estratagema do amor	Marquês de Sade	18114	3806	104
O rei	Isaak Babel	10108	2046	60
Aduzinda e Zulmiro – a magia da adolescência		7055	1453	77

Table. 1 – Collection of stories

3 Solution

First we collected a set of children stories, all of them by Portuguese authors (see Table 1), and divided it as a train set, composed by the first eleven stories, and a test set, composed by the last four stories.

From the analysis of 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 this paragraph 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 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. The 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's the narrator imitating the character and not the character speaking about himself;

Heuristic 8. The tense adverbs produced like in the message (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, so the next excerpt of text probably must belong to a character;

Heuristic 10. The imperfect tense verbs that can be expressed by the same way for a character and for a narrator just lead to a direct discourse when there is a personal pronoun correspondent to a character;

Heuristic 11. In the phrase, if there is an excerpt of text between two dashes where exists a

declarative verb (declare, say, ask, etc.) on third person then we can say that a character expresses the excerpt of text appearing before the left dash;

Heuristic 12. The use of interjections identifies a direct discourse because only the characters use them.

However, when DID was implemented we needed to operate some changes to their use, 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 6** is applied together with **Heuristic 5**, because DID's input has many ambiguities.
- **Heuristic 7** is a neutral heuristic so it's 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 [4], that contains all the identified discourses accordingly to the tags supported by this language. Now, we show a little example of a DID's output document:

```
- <vhml>
  - <references>
      <title>Os três porquinhos</title>
      <url>http : //batatoon . iol .
          pt</url>
    </references>
  - <person name="narrator">
      <p>Era uma vez três porquinhos que
          resolveram sair de casa da mãe
          para irem correr mundo . Um dia
          , ao fim de muitos meses de
          viagem , resolveram construir
          cada um a sua casa . O mais novo
          dos três irmãos , que era muito
          preguiçoso e passava os dias de
          papo para o ar , construiu a
          casa num abrir e fechar de olhos
          : pegou em três ou quatro paus ,
          apanhou um monte de palha que
          havia ali mesmo ao lado e já
          está ! Em menos de meia hora
          tinha a casa feita . Quando
          terminou a tarefa , deitou-se
          regalado à sombra de uma árvore
          ...</p>
    </person>
  - <person name="narrator">
      <p>O porquinho do meio também não
          era assim muito amigo de
          trabalhar ... mas , sempre se
          esforçou m pouco mais que o
          irmão mais novo :</p>
    </person>
  - <person name="" heuristic="h1 , h2 ,
      h4 , h6 , h11">
      <p>- Vou construir uma casa de
          madeira !</p>
    </person>
  - <person name="narrator">
      <p>- disse ele .</p>
    </person>
  - <person name="narrator">
      <p>Serrou algumas tábuas , martelou
          aqui e ali e , no final da manhã
          , tinha pronta a sua cabana
          .</p>
    </person>
  - <person name="narrator">
      <p>O porquinho mais velho trabalhou
          todo o dia : primeiro desenhou a
          casa , calculou bem as medidas ,
          pensou em todos os detalhes .
          Depois construiu a sua casa com
          todos os cuidados : paredes de
          tijolo resistentes , janelas com
          portadas de madeira , tudo
          impecável . Quando acabou o
          trabalho estava muito cansado ,
          mas tinha valido a pena . Ao
          final do dia , estavam os três
          porquinhos contentes na
          brincadeira quando lhes apareceu
          pela frente um lobo esfomeado
          !</p>
    </person>
  - <person name="narrator">
      <p>Os porquinhos correram logo para
          as suas casas . O lobo foi à
          casa de palha :</p>
    </person>
  - <person name="lobo" heuristic="h1 , h2
      , h4">
      <p>- TOC ! TOC ! TOC !</p>
    </person>
  - <person name="" heuristic="h1 , h3">
      <p>- Quem é ?</p>
    </person>
  - <person name="" heuristic="h1 , h4 ,
      h5 , h6">
      <p>- É o lobo . Abre a porta , se
          não eu sopro e faço a casa ir
          pelos ares ! ! !</p>
    </person>
</vhml>
```

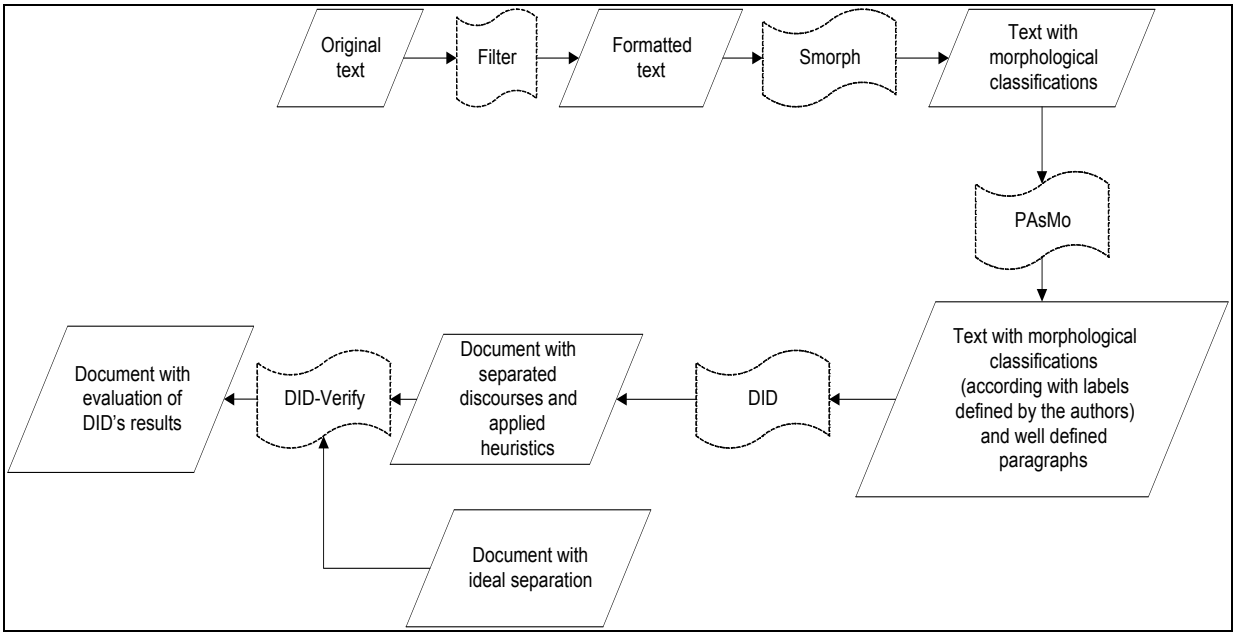

Figure. 2 – Architecture of the DID System.

DID followed the Theory of Confirmation to get the degree of trust with which one direct discourse is identified: the user can define the trust in 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.

Figure 2 contains the information flow between the different modules, which begins with the acquisition of texts and finishes with DID and DID-Verify's output. The Filter is a useful application to transform the input texts syntactically.

4 Discussion

In order to check the capabilities of DID system we developed a new system: the DID-Verify, which is responsible for the comparison between DID's output and one idealized result. This comparison verifies if discourses were well identified by DID and also shows the number of times that each heuristic is applied (see Table. 2).

After analyzing the results obtained with the train 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 "O Bando dos Quatro", 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 to a Story Teller System under development by other researchers of our research institute. The result obtained for the story "Lisboa 2050" 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 what shows that the results are similar to projected objectives.

Story	Correct results	Incorrect results	Success rate
O Gato das Botas	28	0	100%
O Macaco do Rabo Cortado	48	0	100%
O Capuchinho Vermelho	41	1	97%
Os Três Porquinhos	28	1	96%
Lisboa 2050	147	6	96%
A Branca de Neve	43	2	95%
Ideias do Canário	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 (Capítulo 1)	70	40	63%
Pinóquio	43	1	97%
O estratagema do amor	147	11	93%
O rei	81	9	90%
Aduzinda e Zulmiro – a magia da adolescência	95	12	88%

Table. 2 – Results of DID measured by DID-Verify

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

Examining the results obtained by DID-Verify to the test set we designed the Tab. 3, which shows the performance of each applied 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. The Heuristic 4 is the one that leads to more mistakes, because the exclamation mark is many times used in narration discourses.

Heuristic	Number of successes	Number of 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

Table. 3 – Analysis of correctness

Generally, all the heuristics have a high success rate.

5 Future work

We point out the improvements that we plan to introduce in the DID system:

- Define associations of words and expressions to help identify some type of story characters;
- Define a set of verbs that cannot be expressed by a narrator, e.g. *to be* (first person).
- Use a morphossyntatic disambiguator [3] to handle all the ambiguous word classifications present in DID's input.

DID-Names is an interesting challenge as a prolongation of DID system. DID-Names is a system responsible by the identification of the names of each character.
In another phase we say that would be useful a system capable of characterize the characters gesture and emotionally as well as the environment of the story.

However we can say also that these two systems are something complex because they always depend from linguistic characteristics that are sometimes ambiguous or absent.

6 References

[1] A.Silva, M. Vala, and A. Paiva, Papous: The Virtual Storyteller, Intelligent Virtual Agents, 3rd International Workshop, 2001, Madrid, Spain, 171–181, Springer-Verlag LNAI 2190.

[2] Joana 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 2002), 219-227.

[3] R. Ribeiro, L. Oliveira e I. Trancoso, morphossyntatic Disambiguation for TTS Systems, Proceedings of 3rd Conference on Language Resourses and Evaluation (LREC), Las Palmas, Spain, 1427-1431, ELRA.

[4] Gustavson, C., Strindlund, L., Emma, W., Beard, S., Quoc, H., Marriot, A., Stallo, J. – *VHML Working Draft v0.2*, 2001