Character Identification in Children Stories

Nuno Mamede and Pedro Chaleira
{*Nuno.Mamede, Pedro.Chaleira*}*@l2f.inesc-id.pt*

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

Abstract. The automatic identification of direct and indirect discourses and the association of each "direct" utterance with its author are research topics that begin to be explored in Natural Language Processing.
We developed the DID system that when applied to children stories starts by classifying the utterances that belong to the narrator (indirect discourse) and those belonging to the characters taking part in the story (direct discourse). Afterword, DID tries to associate each direct discourse utterance with the character(s) in the story.
This automation can be advantageous, namely when it is necessary to tag the stories that should be handled by an automatic story teller.

1 Introduction

Children stories have some intrinsic magic that captures the attention of any reader. This magic is transmitted by intervening 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 characters and narrator, respectively.

The automatic identification of speakers in children's stories is a necessary step for the comprehension of the story, namely when it is necessary to tag the stories that should be handled by an automatic story teller. After knowing which portions of the story should be read by which speaker, it is possible to choose the appropriate voices for synthesizing the story characters [5], to choose the appropriate character representation and animate it in a story teller [1].

This work deals with the identification of the character (the narrator may be considered another character, for this purpose) that is responsible for each story utterance. The result is expressed in a final document with tags associated with each character.

For example, consider the following excerpt of a story[1]:

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

Our system identifies the text associated with each character of the story:

[1] Although some examples are in english, our system only handles portuguese texts

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

Our approach consists of two basic stages: (i) identification of the utterances that belong to the narrator and the utterances that are said by a character, and (ii) association of each character utterance wit a specific story character. The first stage is described in Section 2 and the second stage is described in Section 3.

2 Direct/Indirect Discourse Identification

2.1 Pre-processing

In order to apply DID it is necessary to resort to Smorph [6], a morphological analyzer, and PasMo [4], 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.

2.2 Solution

We started by collecting a set of children stories, all of them written by Portuguese authors. These dorpora was divided into a training set (the first eleven stories of Table 1), and a test set (the last four stories of the same table). After hand inspecting the training set, we extracted twelve heuristics:

H1 A dash at the beginning of a paragraph identifies a direct discourse;
H2 A paragraph mark after a colon suggests the paragraph corresponds to a character (direct discourse);
H3 If a paragraph has a question mark at the end then it, probably, belongs to a character;
H4 The exclamation mark at the end of a paragraph identifies a direct discourse, with some probability. This heuristic follows the reasoning of H3;
H5 The personal or possessive pronouns in the 1st or 2nd person indicate that we are in the presence of a direct discourse;
H6 Verbs in past tense, present, future or imperfect tense are characteristics of direct discourse because they are verbs directed to characters;
H7 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;
H8 The usage of tense adverbs (tomorrow, today, yesterday, etc.) can identify a direct discourse;
H9 If next to a direct discourse there is a short text between dashes, then the next text excerpt probably belongs to a character;

H10 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;

H11 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;

H12 The use of interjections identifies a direct discourse because only characters use them.

The input of 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 [2], 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.

However, DID first results made us improve these heuristics, namely:

- H3 and H4 have different trust values deppending on the position of the mark on the paragraph. If there is a question or exclamation mark in the middle of a paragraph, the trust value is lower. When the mark is at the end of the paragraph the trust value is higher. So, these heuristics have two trust values.
- H5 and H6 have been combined, because DID's input has many ambiguities.
- H7 revealed to be a neutral heuristic so it was removed.

Final results can be seen in Table 1

Table 1. Performance of direct/indirect discourse separation

Story	Correct	Incorrect	Success rate
1 - O Gato das Botas	28	0	100%
2 - O Macaco do Rabo Cortado	48	0	100%
3 - O Capuchinho Vermelho	41	1	97%
4 - Os Três Porquinhos	28	1	96%
5 - Lisboa 2050	147	6	96%
6 - A Branca de Neve	43	2	95%
7 - Ideias do Canário	41	2	95%
8 - Anita no Hospital	102	11	90%
9 - Os Cinco e as Passagens Secretas	131	19	87%
10 - A Bela e o Monstro	31	6	83%
11 - O Bando dos Quatro: A Torre Maldita (Chap. 1)	70	40	63%
12 - Pinóquio	43	1	97%
13 - O estratagema do amor	147	11	93%
14 - O rei	81	9	90%
15 - Aduzinda e Zulmiro – a magia da adolescência	95	12	88%

2.3 Evaluation

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 "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 for 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 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 the correctness of each heuristic

Heuristic	N Successes	N Failures	Success rate
H1	188	2	98.9%
H2	6	0	100%
H3	59	1	98.3%
H4	37	3	92.5%
H5	81	2	97.6%
H6	70	1	98.6%
H8	7	1	87.5%
H12	17	1	94.4%

3 Character Identification

3.1 VHML Changes

Sometimes, it is not clear, even for humans, which is the character that must be associated with a given utterance. To allow the representation of this king of ambiguity,

and to avoid the repetition of utterances whenever an utterance is simultaneously said by multiple characters, we made small changes to the VHML language. Namely, we introduced the concept of "speaker":

```
<!ELEMENT vhml (paragraph | p | person | references | speaker | mark)+>

<!ELEMENT speaker ((personname|colective)+,person)>
<!ELEMENT colective (personname,personname+)>
<!ELEMENT personname (#PCDATA)>
```

The following example represents an ambiguity: "Text..." must be associated either with a group of characters (`Character1`, `Character2` and `Character3`) or with `Character4`:

```
<speaker>
   <personname>Group1</personname>
   <colective>
      <personname>Character1</personname>
      <personname>Character2</personname>
      <personname>Character3</personname>
   </colective>
   <personname>Character4</personname>
   <person>
      <p>Text...</p>
   <person>
</speaker>
```

3.2 Pre-processing

The text is processed by a shallow parsing module – SuSAna – that performs efficient analysis over unrestricted text. The module recognizes, not only the boundaries, but also the internal structure and syntactic category of syntactic constituents [3]. It is used to identify the nucleus of the noun phrases.

A *single noun phrase* (SNP), is a noun phrase containing either a proper noun or an article followed by a noun phrase.

We *only* considered as declarative the following verbs: *acrescentar*, *acudir*, *adicionar*, *afirmar*, *anunciar*, *aparecer*, *argumentar*, *atalhar*, *atirar*, *avisar*, *chamar*, *comunicar*, *confessar*, *continuar*, *concluir*, *declarar*, *dizer*, *exclamar*, *explicar*, *expor*, *fazer*, *gritar*, *interromper*, *manifestar*, *meter*, *noticiar*, *observar*, *ordenar*, *pensar*, *perguntar*, *publicitar*, *redarguir*, *repetir*, *replicar*, *resmungar*, *responder*, *retorquir*, *rosnar*, *ser*.

The system knows the characters that are referred in each story, which is expressed in a XML file with the following format:

```
<characters>
   <newcharacter>
      <name>Character name</name>
      <gender>male, female or neutral<gender>
      <cardinality>singular or plural</cardinality>
      <alterntivename>
```

```
            <name>Alternative name 1</name>
            <name>Alternative name 2</name>
            <name>Alternative name 3</name>
         </alternativename>
      </newcharater>
      .
      .
      .
   </characters>
```

3.3 Solution

From the hand inspection of the training set (the first eleven stories of Table 1), we extracted five rules. They are not called heuristics to avoid confusions with direct/indirect heuristics, already presented.

Rule 1 If the first sentence of the indirect discourse (imediately following a direct discourse) contains a declarative verb (3rd person) that appears before the first SNP, and the SNP is a valid character name, then that name is the responsible for the previous direct discourse utterance.

Example, from the corpora:
– *Mas não vão conseguir - disse a Ana, inesperadamente.*

Rule 2 If in a direct discourse, any sentence belonging to a previous (search is performed from the direct discourse to the beginning of the document) indirect discourse, the word that precedes a SN containing a character name is a declarative verb on the 3rd person, then the noun of that SN refers the character responsible for the direct discourse.

Example:
E mais uma vez apareceu o lobo mau:
– *TOC! TOC! TOC!*

Rule 3 If in a direct discourse, any sentence belonging to a previous (search is performed from the direct discourse to the beginning of the document) direct discourse starts with a SN containing a character name, then the noun of that SN refers the character responsible for the present direct discourse.

Example:
E eis que a menina bate à porta...
– *Avó, sou eu. Tenho aqui uma pequena prenda para si...*

Rule 4 If the direct discourse itself contains (search is performed from the beginning to the end of the direct discourse) either a SN containing a character name preceded by a verb on present, 1st or 3rd person, or a SN containing a character name imediately preceded by the verb “chamar” in the present, reflexive 1st person, then the noun of that SN refers the character responsible for the present direct discourse.

Example:

Uma rapariga loira e bonita apresentou-se:
– Sou a Sofia e vou tratar da sua viagem. Que filme é que quer ver?

Rule 5 If in a direct discourse, any sentence belonging to a previous (search is performed from the direct discourse to the beginning of the document) direct discourse has a SN containing a character name, not preceded by a declarative verb, or imediately followed by a punctuation mark, then the noun of that SN refers the character responsible for the present direct discourse.

Example:
– Até que enfim que te encontro, Anita! Procurei-te por todo o lado. Tenho uma carta para ti.
– Uma carta? Para mim?

3.4 Evaluation

The evaluation process compares the noun included in the selected NP with the names described in the file that contains the enumeration of all story characters. If the noun is either included (a string operation) in the correct name of the character, or is included in any of the alternate names, we consider it a correct identification. Sometimes, it may lead to a "soft" evaluation, since the name *porquinho* is considered a correct identification of any of the following characters: *porquinho mais velho*, *porquinho do meio*, and *porquinho mais novo*.

The first step was the independent evaluation of each rule, see Table 3. Then we trained a decision tree (CART) to identify the rule that is able to predict the character responsible for each direct utterance. To train the decision tree, we used as features:

1 parameter containing the rule with the correct result;
5 parameters containing the information about the use of each rule: 0 means the rule did not trigger, 1 that the rule did trigger;
10 parameters containing the agreement between rules: 0 means that at least one of the rules did not trigger, 1 the rules agree, and 2 the rules did not agree.

The performance of the decision tree after training was as is depicted in Table 4: 84.8% of correct answers on the training corpus (145 correct answers out of 171), and 65.7% of correct answers on the test corpus (23 correct out of 35). Table 5 contains the confusion matrix after the training.

3.5 Discussion

First of all, the performance achieved is very good, considering that these are the first results, and there is place for improving every stage of the chain processing. One can also conclude that both corpus are small, and the test corpus is not ???????, since it does not contain elements of class 4 (R4 should be applied) and class 5 (R5 should be applied).

The rules, when they decide to trigger, produce good results, but either they should be active more often (68% of direct discourses are not associated with any character),

Table 3. Results of DID measured by DID-Verify

		Rule 1		Rule 2		Rule 3		Rule 4		Rule 5	
Story	**DD #**	**App %**	**Succ %**	**App %**	**Succ %**	**App %**	**Succ %**	**App %**	**Succ %**	**App %**	**Succ %**
1	7	42.8	100	42.8	100	14.3	100	0	—	0	—
2	28	14.3	100	14.3	0	32.1	88.9	0	—	0	—
3	20	5	—	0	—	30	66.7	0	—	0	—
4	15	6.6	100	13.3	100	20	66.7	30	100	0	—
5	91	4.4	100	6.6	50	18.7	52.9	2.2	100	3.3	100
6	13	7.7	100	15.4	50	30.8	75	0	—	0	—
7	36	0	—	0	—	2.8	100	0	—	0.0	100
8	66	7.5	100	9.1	66.7	22.7	40	1.5	100	1.5	100
9	76	46	100	48.7	75.7	3.9	66.7	0	—	1.3	100
10	9	0	—	0	—	22.2	50	0	—	0	—
11	62	6.5	100	8.1	40	8.1	40	0	—	4.8	0.0
Sum	**423**	**13.7**	**98.3**	**15.6**	**65.2**	**15.6**	**59.1**	**1.4**	**100**	**2.1**	**66.7**
12	15	13.3	100	20	33.3	0	—	0	—	0	—
13	104	2.9	100	4.8	80	1	100	0	—	0	—
14	38	5.3	100	5.3	50	7.9	—	0	—	0	—
15	59	6.8	50	6.8	50	23.7	57.1	0	—	0	—
Sum	**216**	**5.1**	**81.8**	**6.5**	**57.1**	**8.3**	**50**	**0**	**—**	**0**	**—**

Table 4. Performance of the Decision Tree on Identifying the Character of each Direct Utterance

	Recall	**Precision**
Training Corpus	34.3%	84.8%
Test Corpus	10.6%	65.7%

Table 5. Confusion Matrix of the Decision Tree on Identifying the Character of each Direct Utterance

	Training corpus							
	–	**R1**	**R2**	**R3**	**R4**	**R5**	**Sum**	**Succ**
–	0	1	11	11	0	2	25	0%
R1	0	57	0	0	0	0	57	100%
R2	0	0	42	0	0	0	42	100%
R3	0	0	1	36	0	0	37	97.3%
R4	0	0	0	0	5	0	5	100%
R5	0	0	0	0	0	5	5	100%
Sum	0	58	54	47	5	7	171	83.3%

	Test corpus							
	–	**R1**	**R2**	**R3**	**R4**	**R5**	**Sum**	**Succ**
–	0	0	5	6	0	0	11	0%
R1	0	9	0	0	0	0	9	100%
R2	0	0	7	0	0	0	7	100%
R3	0	1	0	7	0	0	8	87.5%
R4	0	0	0	0	0	0	0	100%
R5	0	0	0	0	0	0	0	100%
Sum	0	10	12	13	0	0	35	68.6%

or other rules are missing. R4 and R5 should be revised, since they are triggered only 1%.

After a more detailed evaluation of the errors in both corpora we concluded that: (i) only twice a part-of-speech tagger is the culprit; (ii) the incorrect identification of direct/indirect discourse is responsible for 13 mistakes (rule R1-1, R2-2, R3-7 and R5-3); (iii) The shallow parser did not identify a NP and, consequently, R3 failed to identify the correct character 4 times; and the decision tree has made an incorrect choice 24 times.

4 Future work

We have made good progress on children story interpretation/marking, but, since it is a complex and heterogeneous problem, a lot of work remains to be done, namely:

- to increment the size and variety of both corpora;
- to define associations of words and expressions to help identify some type of story characters;
- to use a morphossyntatic disambiguator to handle the ambiguous word classifications;
- to improve the automatic text segmentation in sentences ;
- to define a set of verbs that cannot be expressed by a narrator;
- to identify relations (ownership and creation) between objects and characters;
- to identify the family relations between characters;
- to introduce in the processing chain a new module to identify the characters taking part in a story, which is, for the moment, given by the user;
- to use anaphora to help identifying characters referred by pronouns;
- to implement and evaluate a new rule: whenever two consecutive direct utterances are associated with different characters, and until the next indirect utterance, consider the identified characters alternate in the enclosed direct discourses;
- to include propositional and adjective attachments in the noun phrases, to enable a better identification of the story characters;
- to identify the subject and complements of each sentence, which will enable a better redesign/performance of rules R2 and R3;
- to develop new modules to identify the gesture and emotions as well as the environment where each scene takes place.

Acknowledgments

The Direct/Indirect Discourse Identification task must be credited to Diogo Barbosa and Ricardo Costa, while their stay at L^2F in 2002.

References

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

2. C. Gustavson, L. Strindlund, W. Emma, S. Beard, H. Quoc, A. Marriot, J. Stallo, VHML Working Draft v0.2 (2001).
3. F. Batista, Nuno Mamede, SuSAna: Módulo Multifuncional da Análise Sintáctica de Superfície, in Proc. Multilingual Information Access and Natural Language Processing Workshop (IBERAMIA 2002), Sevilla, Spain, 29-37 (2002).
4. 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).
5. J. Zhang, A. Black, R. Sproat, Identifying Speakers in Children's Stories for Speech Synthesis, Eurospeech 2003, Geneva, Switzerland (2003).
6. S. Ait-Mokhtar: L'analyse présyntaxique en une étape. Thèse de Doctorat, Université Blaise Pascal, GRIL, Clermont-Derrand (1998).