Automatic Aspect Identification: The Case of Informative Microaspects in News Texts

Alessandro Y. Bokan[1 2] and Thiago A. S. Pardo[1 2]

[1] Institute of Mathematical and Computer Sciences, University of São Paulo, SP, Brazil
[2] Interinstitutional Center for Computational Linguistics, São Carlos, SP, Brazil
{abokan,taspardo}@icmc.edu.br

Abstract. Informative aspects represent the basic units of information in texts. For example, in news texts they could represent the following information: what happened, when it happened and where it happened. With the identification of these aspects, it is possible to automate some NLP tasks such as Summarization, Question Answering and Information Extraction. *Microaspects* --a type of informative aspects-- represent local segments of the sentence. In this paper, we automatically identify *microaspects* using Semantic Role Labeling, Named-Entity Recognition, Handcrafted Rules and Machine Learning techniques. We evaluate our proposal using the CSTNews journalistic corpus, which has manually annotated aspects. The results are satisfactory, and prove that *microaspects* can be automatically identified in news texts with acceptable performance.

Keywords: Automatic Summarization, Semantic Role Labeling, Named-Entity Recognition, Machine Learning

1 Introduction

Informative aspects represent semantic-discursive basic units of information present in sentences. Aspects can represent local components of a sentence, like specific location or a certain date. They can also appear from the sentence context. For example, in natural disasters news, the following aspects could be recognized: when it happened?, where it happened?, how it happened?, and what happened?.

Aspects date back to the work of Swales [14] with the model CARS (Create a Research Space), where schematic structures are used to create/organize scientific texts. Recently, the TAC (Text Analysis Conference) – the main conference and scientific competition about Automatic Summarization – proposed the use of informative aspects to assist the Summarization Guided task (2010/2011). For example, TAC analyzed and determined that summaries of the category "Attacks" might include specific aspects as: what happened (WHAT), casualties (WHO_AFFECTED), perpetrators (PERPETRATORS), location (WHERE) and date (WHEN).

Fig. 1 presents an aspect annotation of a summary of the category "Attacks". The first sentence reports that several attacks occurred (WHAT) in São Paulo (WHERE),

on Monday (WHEN). The second sentence identifies the entities (person or organization) affected by the attacks (WHO_AFFECTED). Finally, the last sentence identifies the criminal entities (PERPETRATORS).

[A new series of criminal attacks happened at dawn of Monday, March 7, in São Paulo and near cities.] **WHAT/WHEN/WHERE**
[The bandits attacked banks, police stations and public buildings with bombs and gunshots.] **WHO_AFFECTED**
[Those actions are attributed to the criminal gang "Primeiro Comando da Capital" (PCC), which has led twice others attacks.] **PERPETRATORS**

Fig. 1. Example of aspects annotation

The main goal to identifying aspects is to **automate** some NLP tasks such as Summarization, Question Answering and Information Extraction. According to Genest et al. [7], "*aspect identification can be useful to determine relevant information from source texts and to identify structural constraints to construct summaries"*. Owczarzak and Dang [10] proposed the use of informative aspects as a deep approach to produce coherent and cohesive multi-document summaries for a specific genre (*e.g.*, journalistic, narrative, opinion, etc.) and category (*e.g.*, sports, politics, etc.). Since TAC, aspects have been used in several studies in the literature to assist the summarization task [13,8].

Aspects can indicate standard structures (*templates*) to model criteria of content selection and organization to automatically generate coherent summaries. For that reason, Rassi et al. [12] manually annotated aspects on multi-document summaries in the Brazilian Portuguese journalistic corpus called CSTNews [6].

CSTNews was built mainly to assist the Multi-document Automatic Summarization task. The identification of textual segments in different sentential structural levels to determine informative aspects resulted in the classification of aspects in *microaspects* and *macroaspects*. *Microaspects* represent local segments that make up a sentence. *Macroaspects* emerge from the combination of linguistic patterns contained in the local segments inside a sentence, or from the relationship between two or more sentences.

In this paper we use two approaches to automatically identify *microaspects*. The first approach is based on semantic roles, named-entities and handcrafted rules. The second approach is based on machine learning techniques. We evaluate our proposal using a set of aspect-annotated summaries in the CSTNews corpus. In this work, aspects are specifically defined for **journalistic genre**, based on the TAC's Summarization task.

The remaining of the paper is organized as follows: in Section 2, we introduce some related work; in Section 3, we describe the two approaches used to identify *microaspects*; the experiments and results are presented in Section 4; finally, in Section 5, we conclude this paper.

2 Related Work

2.1 Text Analysis Conference

The "Text Analysis Conference" (TAC) is the main conference and scientific competition about Automatic Summarization (AS). In 2010[1], aspects have been proposed to assist the Guided Summarization task to explore a deeper linguistic analysis of the source documents. The goal was to generate a 100-word summary from a set of 10 "newswire" articles for a given topic. Each topic falls into a predefined category. All the participants in the task were given a list of aspects for each category. Finally, the generated summary should include all aspects defined for its category. Table 1 shows some categories and their defined aspects. The remainder categories are: "Health and Safe", "Endangered Resources" and "Trials and Investigations".

Table 1. Aspects defined for each category

Category	Aspects
Accidents and Natural Disasters	What happened (WHAT); date (WHEN); location (WHERE); rescue efforts (COUNTERMEASURES); damages caused by the accident/disaster (DAMAGES); reasons for accident/disaster (WHY); casualties (WHO_AFFECTED).
Attacks	What happened (WHAT); date (WHEN); location (WHERE); reasons (WHY); casualties (WHO_AFFECTED); entity responsible for the attack (PERPETRATORS); damages caused by the attack (DAMAGES); rescue efforts (COUNTERMEASURES).

Some studies were done using TAC 2010 principles. Steinberger et al. [13] proposed a deep semantic analysis to model informative aspects for multilingual AS. Makino et al. [9] and Li et al. [8] automatically identified informative aspects in Wikipedia and generated summaries based on those aspects. Barrera et al. [3] created a question-answering system, called *SemQuest*, based on aspect identification for different categories. Even before TAC, some works presented similar approaches, for example, White et al. [15] proposed aspects-based templates for summaries of "Natural Disasters" texts.

2.2 CSTNews Corpus

The CSTNews corpus [6] is a resource that contains 50 Brazilian Portuguese journalistic text collections. Each collection has 2-3 documents on the same subject but from different sources. The collections were classified into 6 textual categories: Daily News (14), Sports (10), World (14), Politics (10), Money (1) and Science (1). In addition to the raw texts, CSTNews comprises 140 manually generated single-document abstractive summaries, 50 manually generated multi-document abstractive summaries and 50 manually generated multi-document extractive summaries.

[1] http://www.nist.gov/tac/2010/Summarization/

Rassi et al. [12] annotated aspects over the 50 manually generated multi-document abstractive summaries from CSTNews corpus. CSTNews categories are different from TAC-2010 categories. However, there are similarities between categories, *e.g.*, "Daily news" and "World" could contain "Accidents and natural disasters" topics. As mentioned before, aspects were divided in *microaspects* and *macroaspects*. *Microaspects* represent local segments that make up a sentence. *Macroaspects* emerge from the combination of linguistic patterns contained in the local segments inside a sentence, or from the relationship between two or more sentences. In total, [12] identified 8 *microaspects* (see Table 2). It is important to say that aspects were annotated at the **end of the sentence** (sentential level).

Table 2. CSTNews *microaspects* definition

Microaspect	**Definition**
WHO_AGENT	Entity (person/organization) responsible for the fact/event
WHO_AFFECTED	Entity (person/organization) affected by the fact/event
WHEN	Date/time of occurrence of the fact/event
WHERE	Physical/geographical location of the fact/event
WHY	Why the fact/event happens (reasons)
HOW	How the fact/event occurs
SCORE	Result of the sport event
SITUATION	Situation when the fact/event occurs

Fig. 2 shows an example of an annotated sentence with aspects of the "World" category. Concerning to *macroaspects*, it is reported a natural disaster event (WHAT) and the declaration emitted by the pro-Pyongyang Japanese newspaper (DECLARATION). On the other hand, concerning to *microaspects*, it is reported that the disaster happened in July (WHEN), in North Korea (WHERE), because of the floods (WHY) and leaving many dead and injured people (WHO_AFFECTED).

[A study from the japanese newspaper that supports pro-Pyongyang says that, in decorrence of the floods that hit North Korea in july, at least 549 people passed away, 3.043 people were wounded and 295 people are missing.]**WHAT,WHEN, WHO_AFFECTED,WHY,WHERE,DECLARATION**

Fig. 2. Annotated sentence of a summary in the CSTNews corpus

2.3 PALAVRAS Parser

PALAVRAS is a rule-based syntactic parser for Portuguese developed by [4]. In addition, it produces a list of semantic tags[2]. It has two output formats: a simple format ("flat"), and a traditional syntactic tree format ("tree").

According to its author, PALAVRAS achieved a correctness rate of over 99% for morphology and part-of-speech. For syntax the figures are 97-98%. In this work, we

[2] http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#semtags_nouns

use PALAVRAS as input for the semantic role classifier (*tree* format) and to create an aspect classifier using Machine Learning techniques (*flat* format).

2.4 Semantic Role Labeling

Semantic roles represent the semantic relationships between verbs and their arguments. The task of identifying which phrases act as arguments of a particular verb is called Semantic Role Labeling (SRL). For Brazilian Portuguese, [1] proposed a supervised classification system that consists of 3 phases: (1) target verb identification, (2) argument identification and (3) argument classification.

Fig. 3 shows an example of the SRL process. Firstly, the target verb "won" is identified. Secondly, arguments (*A*) are identified: "Brazilian team", "Finnish team" and "in Tampere". Lastly, arguments are annotated with semantic roles: A0 (agent), A1 (patient) and AM-LOC (location), respectively. The "A" terminology refers to the "argument" followed by a prototypical number (1-5), and the "AM" terminology refers to the "modifier argument", followed by a type of modifier, such as time, location, cause, etc. Semantic roles were defined by [11]. It achieved a F1 measure of 94.5% in the identification phase and 81.70% in the classification phase, being the best system for Brazilian Portuguese.

The Brazilian team [*won*] over the Finnish team in Tampere. (1)
[The Brazilian team]$_{A}$ [*won*] [over the Finnish team]$_{A}$ [in Tampere]$_{A}$. (2)
[The Brazilian team]$_{A0}$ [*won*] [over the Finnish team]$_{A1}$ [in Tampere]$_{AM\text{-}LOC}$. (3)

Fig. 3. Semantic Role Labeling annotation example

Arguments related to the verb can answer some questions like who, where, when, why and how. In the previous example, the answers to the questions who won?, who lost? and where won?, are "Brazilian team", "Finland team" and "Tampere", respectively. Therefore, semantic roles can define informative aspects such as WHO_AGENT (who won?), WHO_AFFECTED (who lost?) and WHERE ("where won?). In this paper we propose the use of the SRL system for Brazilian Portuguese developed by [1] to identify some *microaspects*. Table 3 presents the proposed equivalences among some *microaspects* and the corresponding semantic roles.

Table 3. Equivalences among *microaspects* and semantic roles.

Microaspect	**Semantic role**	**Name**	**Definition**
WHO_AGENT	A0	Agent	Subject who did the action
WHO_AFFECTED	A1	Patient	Subject affected by the action
WHERE	AM-LOC	Location	Where the action happens
WHEN	AM-TMP	Temporal	When the action happens
HOW	AM-MNR	Manner	How the action was performed
WHY	AM-CAU	Cause	Reasons for the action

2.5 Named-entity Recognition (NER)

Named-entities (NE) are concrete or abstract entities referenced in the text by a proper noun. Named-Entity Recognition (NER) is a subtask of Information Extraction that aims to identify and classify text entities of predefined categories such as LOCATION, TIME, and EVENT, among other categories of interest.

HAREM[3] is an evaluation event of NER systems for Portuguese document collections. Many works were presented in the two editions of the HAREM. One of the best open-source systems, presented in the second HAREM, was REMBRANDT [5]. REMBRANDT intensely explores Wikipedia as a knowledge source and executes a set of grammatical rules to take advantage of the internal and external indications of the NEs to extract its significance. Furthermore, REMBRANDT has a proper interface to interact with Wikipedia, SASKIA, that provides simple category navigation. That system has 56.74% F1 measure for NER.

NE categories can define some informative aspects: WHERE is equivalent to LOCATION, WHEN to TIME, and SITUATION to EVENT. Thus, we propose the use of REMBRANDT system to automatically identify some *microaspects*. Table 4 shows the proposed equivalences among *microaspects* and NE categories.

Table 4. Equivalences among *microaspects* and NE categories

Microaspect	NE Category
WHERE	LOCATION
WHEN	TIME
SITUATION	EVENT

3 Methodology

The automatic identification process was divided in three phases: (1) to compile all the 322 sentences of the 48 CSTNews annotated summaries of categories Daily, Sports, World and Politics; (2) to automatically annotate sentences with *microaspects* using the 3 proposed systems, called SRL, SRL+Rules, REMBRANDT; and Machine Learning techniques; (3) finally, to get a set of annotated sentences.

3.1 SRL System

The SRL system automatically annotates *microaspects* equivalent to semantic roles (*cf.* Table 3). It covers WHO_AGENT, WHO_AFFECTED, WHEN, WHERE, WHY, and HOW. Given a set of sentences, we use PALAVRAS parser *tree* format to generate syntactic trees for each sentence. Trees are represented in *TigerXML* format. Auxiliary verbs were not considered. Following, each instance was annotated by the Alva-Manchego [1] classifier. Semantic roles are then mapped to the corresponding *microaspects* (see Fig. 4). Finally, all annotated aspects are positioned at the **end** of the sentence.

[3] http://www.linguateca.pt/harem/

<aspect SRL="WHO_AGENT"> The Brazilian team </aspect> **won** <aspect SRL="WHO_AFFECTED"> over the Finnish team </aspect> <aspect SRL= "WHERE"> in Tampere </aspect>.

Fig. 4. *Microaspect* sentence annotation using SRL system

3.2 SRL+Rules System

The SRL+Rules system uses handcrafted rules based on patterns founded in **false negatives** and **false positives** of the SRL system, in order to improve its performance. Rules were created for aspects WHO_AGENT, WHO_AFFECTED, WHEN, WHERE and WHY. Also, SCORE rules were defined, despite not existing equivalence with any semantic role. It is crucial to specify that all rules were **specifically created** for Brazilian Portuguese. There is a language dependency.

WHO_AGENT and WHO_AFFECTED rules are only based on false positives, because the SRL system does not identify if an annotated segment represents an entity, person or organization (see Fig. 5). Based on the PALAVRAS's semantic tags, it is possible to determine if an SRL annotated segment has at less a token that represents an entity, PERSON or ORGANIZATION.

Rules for aspects WHEN, WHERE and WHY are based on false negatives and false positives. WHEN rules follow Baptista et al. [2] theory to identify temporal expressions (see Fig. 6). WHERE rules identify local expressions (see Fig. 7) and WHY rules identify causative expressions (see Fig. 8).

Finally, SCORE rules were created based on a few number of annotated sentences (10) and integrated to de SRL+Rules system. Thus, SCORE rules are limited to this minimum set of annotates sentences.

PERSON[4] = [H, HH, Hattr, Hbio, Hfam, Hideo, Hmyth, Hnat, Hprof, Hsick, Htit, hum]
ORGANIZATION = [admin, org, inst, media, party, suborg]
$\in$ = "is an element of"
$\notin$ = "is not an element of"

Rule 1: If a sentence has a segment annotated by the SRL system that contains a token associated to PALAVRAS's semantic tags PERSON or ORGANIZATION, and iff the token is not a REPENTINO's[5] local lexicon, then the segment will be correctly annotated as WHO_AGENT or WHO_AFFECTED. Otherwise, the annotation will be removed.

Input: *<aspect SRL=WHO_AGENT>The president</aspect> says that some constructions are already underway, and <aspect SRL="WHO_AGENT">others</aspect> will start soon.*

president_(Hprof) $\in$ PERSON $\notin$ local_lexicon

others(diff) $\notin$ PERSON

Output: *<aspect SRL=WHO_AGENT>The president</aspect> says that some constructions are already underway, and others will start soon logo.*

Fig. 5. WHO_AGENT/WHO_AFFECTED rules

[4] PALAVRAS's semantic tags for entities PERSON and ORGANIZATION.
[5] Named-entity lexicon created in HAREM-2005 (www.linguateca.pt/repentino/).

PREP = [de, em, a, por, para]
PRON = [ele(s), ela(s), este(s), esta(s), esse(s), essa(s), aquele(s), isto, isso, aquilo, aqui, aí, ali, outra(s)]
DET = [a(s), o(s), um, uns, uma, umas, à(s)]
day_lexicon = [segunda-feira, terça-feira, quarta-feira, quinta-feira, sexta-feira, sábado, domingo]
time_adverb_lexicon = [hoje, amanhã, ontem, anteontem, tarde, madrugada, noite, meia-noite, manhã...]
time_lexicon = [microssegundo, segundo, minuto, hora, dia, semana, mês, ano, década, milênio, época...]
"+/-" = follow_or_not

Rule 1: If a sentence has PREP + (PRON|DET) + time_adverb_lexicon + PREP + (PRON|DET) + day_lexicon +/- NUM, then the sentence is annotated as WHEN.

"A chuva complicava o trânsito na manhã desta segunda-feira, 16."

na_(PREP+PRON) + manhã_(time_adverb_lexicon) + desta_(PREP+PRON) + segunda-feira_(day_lexicon) + NUM

Rule 2: If a sentence has PREP + (PRON|DET) + day_lexicon, then the sentence is annotated as WHEN.

"Um terremoto atingiu Japão nesta segunda-feira matando 9 pessoas."

nesta_(PREP+PRON) + segunda-feira_(day_lexicon)

Rule 3: If a sentence has PREP + (PRON|DET) +/- (TOKEN|NUM) + time_lexicon, then the sentence is annotated as WHEN.

"Aos 18 minutos, Maicon fez o primeiro gol."

Aos_(PREP+DET) + 18_(NUM) + minutos_(time_lexicon)

Rule 4: If a sentence has PREP + (PRON|ARTG) + time_adverb_lexicon, then the sentence is annotated as WHEN.

"A quarta medida foi aprovada nesta madrugada."

nessa_(PREP+PRON) + madrugada_(time_adverb_lexicon)

Fig. 6. WHEN rules

Rule 1: If a sentence has a segment annotated by the SRL system that contains a PREPOSITION "em", followed or not by the DETERMINER/PRONOUN, followed by a NOUN other than a temporal expression, then the segment will be correctly annotated as WHERE. Otherwise, the annotation will be removed.

Input: *"<aspect SRL=WHERE>On Sunday</aspect>, a bloody battle took place."*

on_(PREPOSITION) + Sunday_(temporal expression)

Output: *"On Sunday, a bloody battle took place."*

Rule 2: If a sentence has "em" + uppercase expression, then the sentence is annotated as WHERE.

"O senador Marcos nasceu em São Paulo"

em_(PREPOSITION) + São Paulo_(uppercase expression)

Fig. 7. WHERE rules

cause_lexicon = [por isso, com isso, porque, devido a, por causa de, por força de, em função de, em virtude de, em razão de, em decorrência de, em consequência de, pois, visto que, já que, causado]

Rule 1: If a sentence has cause_lexicon, then the sentence is annotated as WHY.

"O senador teve seu estado de saúde piorado, por causa de complicações gastrointestinais."

por causa de_(cause_lexicon)

Rule 2: If a sentence has the PREPOSITION "por" + infinitive_verb, then sentence is annotated as WHY.

"Já Poliana Okimoto ficará fora de a decisão de os 800m livre por estar com infecção intestinal."

por_(PREPOSITION) + estar_(infinitive_verb)

Rule 3: If a sentence has expression "graças a" + DETERMINER, without being part of "dar graças" expression, then the sentence is annotated as WHY.

"Graças ao médico, o paciente sobreviveu."

Graças_a_(expression) + o_(DETERMINER)

Fig. 8. WHY rules

3.3 REMBRANDT System

The REMBRANDT system automatically annotates *microaspects* equivalent to the NE categories (*cf.* Table 4). It covers WHEN, WHERE, and SITUATION. Fig. 9 shows a sentence annotated using the REMBRANDT system. Note that the segment "Pan American Games" represents the entity EVENT, "Tuesday" represents the entity TIME, and "Maracazinho" represents the entity LOCAL.

The Brazilian volleyball team won over the Finnish team this <aspect EM="WHEN">Tuesday</aspects> by 3 sets to 0 in <aspect EM="WHERE"> Maracazinho</aspect>, on the <aspect EM="SITUATION">Pan American Games</aspect>.

Fig. 9. *Microaspect* sentence annotation using REMBRANDT system

3.4 Machine Learning Approach

Nowadays, computers have the capacity to automatically learn tasks based on experiences. These experiences are formed by a set of examples called "instances". In this work, the task to be learned is the "microaspect identification". With a manually annotated corpus the task follows a Machine Learning (ML) supervised paradigm, where the training set is formed by *instance-class* pairs called *labeled data*. Thus, the instances are represented by the sentences in the corpus, and the classes are represented by the annotated aspects in the sentences. For that reason, we proposed to use ML techniques to create a *microaspect* classifier.

Microaspect identification is a multi-label classification problem. In this work we apply the "problem transformation methods", which aims to transform the multi-label classification problem into a set of binary classification problems. Therefore, many classifiers were created. We only chose the best classifier for each *microaspect*. In total we obtain 8 binary classifiers (see Table 2).

In total, six types of features provided by PALAVRAS *flat* format were defined: bag of words, lemmas, POS (part-of-speech), semantic tags, lemmas+POS and POS+semantic-tags. Each feature is represented by unigrams "(1, 1)", bigrams "(2, 2)" and bigrams+trigrams "(2, 3)". The result of the representation of the six types of features generates many classifiers for each *microaspect*. For example, the classifier "(2, 3) POS" was created based on all bigrams and trigrams of the part-of-speech of all the words in the corpus. The total number of created classifiers is 144.

We use the SVM (Support Vector Machine) supervised algorithm to classify *microaspects* in a sentential level. This algorithm is the most used in NLP classification tasks because, it is the best to deal with big dimensional space vectors.

4 Experiments and Results

The evaluation was measured by 4 metrics: R (Recall) – percentage of actually positive instances that were labeled as such; P (Precision) – percentage of instances labeled as positive that actually belong to this class; F1 (F1-score) – harmonic mean of the P and R; A (Accuracy) – total number of hits over the total number of instances.

All systems were tested on a set of sentences manually annotated with aspects from the CSTNews corpus. In addition, we tested the combination of the systems (SRL+REMBRANDT, SRL+Rules+REMBRANDT) only on the aspects WHEN and WHERE. In total, there are 322 annotated sentences. Table 5 shows the results of the best systems. The best result was obtained by the SRL+Rules for SCORE (F1=1.000) for the class "YES", whereas all the SCORE rules were created only for a few numbers of false positive sentences. In a majority of cases, the best results were obtained by SRL+Rules system. That proves that the handcrafted rules improved the performance of the SRL system. The worst result was found in the SRL system for HOW (F1=0.040). That happens because the SRL system considerably failed (many incorrect annotated sentences), and, in some cases, human annotators failed. REMBRANDT is the only system that can identify SITUATION. Note that all results of both, F1 of the class "NO" and the accuracy (A), are the highest.

Table 5. Best results using system approach

		"YES" class			**"NO" class**			**A**
Microaspect	**System**	**R**	**P**	**F1**	**R**	**P**	**F1**	
WHO_AGENT	SRL+Rules	0.592	0.664	0.626	0.797	0.743	0.769	0.624
WHO_AFFECTED	SRL+Rules	0.417	0.368	0.391	0.836	0.862	0.849	0.758
WHEN	SRL+Rules	0.947	0.504	0.657	0.717	0.978	0.827	0.770
WHERE	SRL+Rules	0.804	0.474	0.596	0.812	0.952	0.876	0.811
WHY	SRL+Rules	0.469	0.789	0.588	0.986	0.944	0.966	0.935
HOW	SRL	0.111	0.024	0.040	0.872	0.972	0.919	0.851
SITUATION	REMBRANDT	0.231	0.750	0.353	0.993	0.933	0.962	0.929
SCORE	SRL+Rules	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Differently from the previous systems, the ML classifier was trained and tested with all the 322 sentences. We used a **stratified** strategy on the corpus to ensure the same proportion of classes in each subset. The corpus was ten times stratified, where the sentences were divided into 70% for training (225 instances) and 30% for testing (97 instances), for each iteration. Thus, the final result is the average value of the iterations. Table 6 shows the best classifier for each *microaspect*. The best result was obtained by the "(1, 1) semantic" classifier for WHEN (F1=0.615). That occurs due to the fact that the PALAVRAS semantic tags contain a time lexicon. The worst result was obtained for microaspect SCORE (F1=0.000). This happens because CSTNews has few sentences annotated with SCORE. Note that most of the best classifiers are represented by unigrams "(1, 1)". Finally, in the majority of the cases the classifier "(2, 3) POS+semantic" obtained the best results because it have more linguistic knowledge (part-of-speech and semantic) than the others classifiers.

5 Conclusions

In this work we proposed a method to automatically identify *microaspects*. *Microaspects* represent local segments that form a sentence. Firstly, we evaluate a system approach based on semantic roles, named-entity categories, handcrafted rules and a

combination of them, using de CSTNews corpus. The results proves that the SRL+Rules system obtained the best result in the majority of the cases. That means that handcrafted rules improved the SRL system performance. However, there were some problems in the identification process: the SRL system failed to correctly identify some sentences, affecting the performance of SRL+Rules system, and the PALAVRAS parser failed to analyze some sentences.

Table 6. Best classifiers using the Machine Learning approach

Microaspect	**Classifier**	**R**	**P**	**F1**	**A**
WHO_AGENT	(2,3) POS+semantic	0.538	0.636	0.583	0.691
WHO_AFFECTED	(1, 1) lemmas	0.222	1.000	0.364	0.854
WHEN	(1, 1) semantic	0.522	0.750	0.615	0.845
WHERE	(2, 3) POS+semantic	0.471	0.615	0.533	0.856
WHY	(2, 3) POS+semantic	0.200	0.500	0.286	0.897
HOW	(1, 1) bag_of_words	0.250	1.000	0.400	0.938
SITUATION	(1, 1) lemmas+POS	0.333	1.000	0.500	0.959
SCORE	All	0.000	0.000	0.000	0.000

Secondly, we evaluate a ML approach based on lexical, part-of-speech and semantic features. The results are not satisfactory because of the few number of annotated sentences. We believe that ML results could be better with more annotated instances in the corpus.

We can not compare the two approaches, because both approaches (system and ML) were tested in different ways. It is important to say that the subjectivity present in the process of the corpus manual annotation could affect the performance of the two approaches.

In conclusion, the system approach proves that *microaspects* can be automatically identified for Portuguese news texts. The main goal to identify aspects is to automate the Summarization task and to assist other NLP tasks (e.g., Question-Answering). This is a *novel* work for Brazilian Portuguese. Finally, we created a baseline to compare results with future systems, e.g. bag of words (unigramas).

Future work will focus on creating a *macroaspect* classifier. Our aim is to build automatic classifiers that cover all aspects defined in the CSTNews corpus (previously defined by TAC).

Acknowledgments. Part of the results presented in this paper were obtained through research on a project titled “Semantic Processing of Texts in Brazilian Portuguese”, sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

References

1. Alva-Manchego, F.: Anotação Automática Semissupervisionada de Papéis Smânticos para o Português do Brasil. Dissertação, Instituto de Ciências Matemáticas e de Computação - ICMC-USP (2013)

2. Baptista, J., Hagège, C. and Mamede, N: Capítulo 2: Identificação, Classificação e Normalização de Expressões Temporais do Português: A Experiência do Segundo HAREM e o Futuro. In: C. Mota e D. Santos (eds.), Desafios Na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM. Linguateca (2008)
3. Barrera, A., Verma, R., and Vincent, R.: SemQuest: University of Houston's Semantics-based Question Answering System. In: Proceedings of the Fourth Text Analysis Conference, Maryland, USA. National Institute of Standards and Technology. (2011)
4. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Ph.D. thesis, University of Arhus. Denmark (2000)
5. Cardoso N: REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In: C. Mota & D. Santos (eds.), Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM. Linguateca, 2008, pp. 195-211 (2008)
6. Cardoso, P., Maziero, E., Castro Jorge, M., Seno, E., Di Felippo, A., Rino, L., Nunes, M., and Pardo, T.: A Discourse Annotated Corpus for Single and Multi-document Summarization of News Texts in Brazilian Portuguese. In: Proceedings of the 3rd RST. Brazilian Meeting, pp. 88-105 (2011)
7. Genest, P., Lapalme, G., e Yousfi-Monod, M.: HEXTAC: the Creation of a Manual Extractive Run. In: Proceedings of the Second Text Analysis Conference, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (2009)
8. Li, P., Wang, Y., Gao, W., and Jiang, J.: Generating Aspect-oriented Multidocument Summarization with Event-aspect Model. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pp. 1137-1146, Stroudsburg, PA, USA. Association for Computational Linguistics (2011)
9. Makino, T., Takamura, H., and Okumura, M.: Balanced Coverage of Aspects for Text Summarization. In: Proceedings of the Fourth Text Analysis Conference, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (2011)
10. Owczarzak, K. and Dang, H.: Who wrote What Where: Analyzing the Content of Human and Automatic Summaries. In: Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, pp. 25-32, Portland, Oregon. Association for Computational Linguistics (2011)
11. Palmer, M., Gildea, D., e Xue, N.: Semantic Role Labeling. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers (2010)
12. Rassi, A., Zacarias, A., Maziero, E., Souza, J., Castro, L., Balage, P., Cardoso, P., Camargo, R., Agostini, V., Filippo, A., Seno, E., Rino, L., and Pardo, T.: Anotação de Aspectos Textuais em Sumários do Córpus CSTNews. Relatório Técnico NILC TR-13-01, Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (2013)
13. Steinberger, J., Tanev, H., Kabadjov, M., and Steinberger, R.: JRC's Participation in the Guided Summarization Task at TAC 2010. In Proceedings of the Text Analysis Conference, pp. 1-12 (2010)
14. Swales, J. Genre Analysis: English in Academic and Research Settings. Cambridge University Press, Cambridge, UK (1999)
15. White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D., and Wagsta, K.: Multidocument Summarization via Information Extraction. In: Proceedings of the First International Conference on Human Language Technology Research, HLT '01, pp. 1-7, Stroudsburg, PA, USA. Association for Computational Linguistics (2001)