Capítulo 3

MUC vs HAREM: a contrastive perspective

Nuno Seco

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 3, p. 35–41, 2007.

This chapter presents a brief overview of two pioneering evaluation contests in the field of Named Entity Recognition (NER) and delves into the conceptual underpinnings of each. The intention is not one of divulging the referred events, as that has been done in Grishman e Sundheim (1996) and Santos et al. (2006), but rather one of contrastive scrutiny. The reader should be attentive of the fact that I am comparing two events that took place in completely different time frames. Notwithstanding, this comparison is relevant because both correspond to the genesis of the joint evaluation paradigm in the field of NER of two different languages, English and Portuguese, respectively.

The field of Natural Language Processing (NLP) has faced many obstacles since its birth. While some have been somewhat overcome, others still remain. One such obstacle is the **identification** and **classification** of named entities. It is in the classification facet of named entities that HAREM differs quite significantly from the Message Understanding Conferences (MUC) (Sundheim, 1995; Grishman e Sundheim, 1996; Hirschman, 1998). Nonetheless, there are evolutions of MUC contests (namely the Automatic Content Extraction (ACE) (Doddington et al., 2004) that address some of the shortcomings pointed out in this chapter. Arguably, many may refute the relevance of this paper because of the time gap between the two events; even so, the discussion is still appropriate as both correspond to the origins of the evaluation event in each language.

The reader should also take into account the fact that by comparing two evaluation events pertaining to two different languages certainly raises issues of authority of such comparisons as is pointed out in Cardoso (2006a, Section 5.3.3). Nonetheless, my concern is not one of comparing the results of the events but one of comparing the underlying assumptions and motivations of these events.

The rest of this chapter is organized in the following manner: Section 3.1 provides a brief overview of MUC, focusing on the aspects dealing with NER. Section 3.2 presents HAREM, contrasting it with MUC along with its guiding principles that motivated the construction of a new evaluation methodology. Section 3.3 presents the fine grained evaluation metrics employed along with their possible combinations. Finally, Section 3.4 concludes the paper summarizing the main differences identified.

3.1 An Overview of MUC

Prior to MUC, several Information Extraction (IE) systems were developed, but most of them were developed having specific domains in mind. Consequently, it was impossible to compare systems and strategies in a just way. As such, the need for a common evaluation environment that would enable fair comparison of systems was acknowledged. In order to quench the need, an informal poll of NLP groups was carried out to determine which groups had running text processing systems, and whether these groups would be interested in coming together to assess the state of NLP systems.

The first MUC event took place in 1986 (Grishman e Sundheim, 1996) and had the main goal of gathering researchers interested in the topic of IE. For the first time, a common corpus of real messages was used, from a common theme (the naval domain). System performance was compared using this common corpus, and the output of each system was discussed.

In 1989 a second MUC event took place and introduced the notion of template filling along with several evaluation metrics. In this edition, the participants had to fill templates that had several types of attributes with their corresponding values extracted from the given text. Introducing such templates and manually pre-calculating the correct values allowed, for the first time, the use of evaluation metrics such as precision, recall or F-measure to measure and compare the system's performances.

From 1991 up to 1993, MUC organized three more evaluation events. The main characteristics of these events was the change in target domains, the size of the corpus, the complexity of the templates and, finally, the inclusion of more languages such as Japanese, Spanish and Chinese.

MUC-6 took place in 1995 and had 3 main goals in mind:

1. Promote the development of reusable components that could be easily used in other NLP related tasks besides IE.
2. Promote, as much as possible, an effortless portability of systems to other domains for which they were not initially conceived.
3. Look into issues concerned with deeper understanding of the texts, such as anaphoric references and relations between attributes of different templates.

Thus, it was in the context of MUC-6 guidelines that NER was identified as being an autonomous component prone task and received diligent attention. MUC-7 took place in 1998 and did not diverge when compared to its preceding event, being that the basic difference was in the number of texts used in the contest.

3.2 Named Entity Recognition

Named entities, from a MUC viewpoint, were defined as: (Sundheim, 1995)

> *"... markables [named entities] includes names of organizations, persons, and locations, and direct mentions of dates, times, currency values and percentages. Non-markables include names of products and other miscellaneous names ('Macintosh', 'Wall Street Journal', 'Dow Jones Industrial Average') ..."*

This definition alone represents a major difference between HAREM and MUC, a discussion postponed to Section 3.3.

NER is considered to be domain independent and task independent, according to MUC's guidelines. The results obtained in MUC's NER task seem to suggest that NER is an easy task, with more than half of the systems obtaining results above 90% in terms of precision and recall (the best system obtained an F-measure of 0.9642).

Before accepting that the NER task is a solved case, one should address the issue of what exactly is being evaluated: The MUC-6 NER task used a golden collection of 30 articles taken from the Wall Street Journal (WSJ) from January of 1993 to June of 1994. MUC-7 used 100 articles from same collection. The named entities of this golden collection were manually identified and classified according to three different categories and subtypes (Sundheim, 1995):

1. `ENAMEX` – Entity names with subtypes organization, people and location.
2. `TIMEX` – Temporal expressions with subtypes date and time.
3. `NUMEX` – Numeric expressions with subtypes money and percent.

Summing up, the classification facet of NER in MUC evaluations was done according to the above mentioned categories. The next section discusses the HAREM evaluation and delineate the underlying conceptual differences in the evaluation.

3.3 HAREM

In HAREM, the classification system of MUC-6 was challenged, questioning its appropriateness to real applications, and if it really represents the NER issue. Note that the categories chosen for MUC were accomplished in a top down manner. On the contrary, HAREM took a bottom-up approach by manually analyzing text, identifying relevant entities and then attributing them a classification in context. As a consequence, a much finer grained classification hierarchy with 10 categories and 41 types was established (Santos e Cardoso, 2006):

1. **`PESSOA`**`:INDIVIDUAL, CARGO, GRUPOIND, GRUPOMEMBRO, MEMBRO, GRUPOCARGO`
2. **`ORGANIZACAO`**`:ADMINISTRACAO, EMPRESA, INSTITUICAO, SUB`
3. **`TEMPO`**`:DATA, HORA, PERIODO, CICLICO`
4. **`LOCAL`**`:CORREIO, ADMINISTRATIVO, GEOGRAFICO, VIRTUAL, ALARGADO`
5. **`OBRA`**`:PRODUTO, REPRODUZIDA, PUBLICACAO, ARTE`
6. **`ACONTECIMENTO`**`:EFERMIDE, ORGANIZADO, EVENTO`
7. **`ABSTRACCAO`**`:DISCIPLINA, ESTADO, ESCOLA, MARCA, PLANO, IDEIA, NOME, OBRA`

8. **COISA**:CLASSE, SUBSTANCIA, OBJECTO, MEMBROCLASSE
9. **VALOR**:CLASSIFICACAO, QUANTIDADE, MOEDA
10. **VARIADO**:OUTRO

Note: COISA:MEMBROCLASSE appeared only on 2006 event. In 2005, OBRA:PRODUTO was discarded.

These finer grained categories lead to a finer grained NER classification task, therefore making the HAREM NER task much more intricate when compared to MUC's task and of other events. Another important aspect that HAREM took into account was **context**, that is, the surroundings in which a named entity appears determines its meaning and, therefore, its category (or categories). For example, in MUC the term **Brasil** would be considered an ENAMEX regardless of the context it appeared in. On the other hand, HAREM dealt with the issue of sense extensions such as metonymy. Consequently, the term **Brasil** could be classified differently according to the surrounding context. Consider the following examples taken from Santos (2006a):

*O **Brasil** venceu a copa...* (PESSOA:GRUPOMEMBRO)
*O **Brasil** assinou o tratado...* (ORGANIZACAO:ADMINISTRACAO)
*O **Brasil** tem muitos rios...* (LOCAL:ADMINISTRATIVO)

In each example, the same term is classified according to the context it appears, an aspect not dealt by MUC. Nonetheless, ACE, for instance, takes this aspect into consideration (Doddington et al., 2004).

Another aspect, and probably the most distinctive aspect is that HAREM, takes vagueness into account during identification and classification. That is, the possibility of a named entity simultaneously being identified or interpreted according to different referents both of which are correct. The issue of vagueness is more carefully discussed in Chapter 4. Consider the following example:

*...era um teólogo seguidor de **Emmanuel Swendenborg**.*
(PESSOA:INDIVIDUAL or ABSTRACCAO:OBRA ?)

In this example, both interpretations are equally acceptable (the writings of the person or the actual person), and most probably they occur simultaneous in our conceptual system and discourse structure (Pustejovsky, 1994). For an in-depth discussion on vagueness in the realm of HAREM we refer the reader to Santos e Cardoso (2006). Nonetheless, MUC also allowed alternative identifications through the use of the ALT tag attribute, but regarding semantic classification was more conservative. For example, the MUC guidelines state that *the White House* should be marked up as ORGANIZATION or have no markup at all in the answer key. This is a highly conservative approach when compared to HAREM that allowed different categories to occur simultaneously.

3.4 Evaluation

In HAREM, a golden collection of 129 (and later another set of 128 different texts for the Mini-HAREM[1] event) texts manually tagged was used as the reference for evaluation purposes. The collection comprised several different text genres written according to several different language varieties, mainly from Portugal, and Brazil, but also from Angola, Mozambique, East Timor, Cape Verde, India and Macao. As well as identifying and semantically classifying named entities, HAREM took into consideration the gender and number of the entities, introducing two new facets of evaluation with subtypes. HAREM proposed 3 subtasks: Identification (correct delimitation of the entity), Semantic Classification and Morphological Classification (gender and number).

Each of these dimensions was evaluated using different configuration scenarios. These have been clearly explained in Chapter 18 and as such it will suffice to say that there are 12 different possible evaluation scenarios for the participant. The motivation for such flexibility is that many participants are only concerned with certain aspects of classification (e.g. only interested in the PESSOA category).

Another issue worth stressing is that the HAREM evaluation software deals with partial alignments. In other words, it can cope with inexact matches of named entities between source and target texts. This aspect was never considered in other evaluation events. A finer discussion of the evaluation aspects of HAREM may be found in Seco et al. (2006).

The metrics used in HAREM subsume the ones proposed and employed in MUC, HAREM introduced many new evaluation metrics (Cardoso, 2006a). Nonetheless, regarding the metrics that were employed in both, the results obtained were drastically different. The best system in the first HAREM event attained an F-measure of 0.63 (considering an evaluation configuration equivalent to that of MUC). At first sight this seems to indicate that the state of the art of NER for Portuguese is substantially inferior to that of English. But from another standpoint one may argue that it is not the quality of NER systems that is inferior to that of English, but that the evaluation standards are much more meticulous in HAREM, resulting in a more demanding task and yielding lower performance values. It is the author's belief that the last perspective correctly mirrors the reality of HAREM.

3.5 Final Remarks

In conclusion, HAREM has brought significant contributions to the field of NER, specifically regarding the Portuguese language, where previous work did not exist. A finer grained classification system has been proposed that was obtained using bottom-up analysis approach of actual corpora. Named entities were classified in context according the classification system proposed; the number of different interpretations in HAREM was con-

[1] The interested reader should see Cardoso (2006a) for details.

siderably larger than in MUC (see Chapter 4). Vagueness, a ubiquitous characteristic of language, was taken into account in the HAREM evaluation. Morphological classification (gender and number) was also considered for the first time in the field of NER. The golden collection employed and used in the evaluation process was substantially wider-ranging when compared to MUC. MUC used the Wall Street Journal, which can be considered a domain specific journal, while HAREM used documents from general newspapers in Portugal and Brazil, Web texts, literary fiction, transcribed oral interviews and technical text. Finally, the evaluation framework showed to be very powerful, fulfilling the assorted needs of the several participants in a very flexible manner.

Acknowledgements

I would like to thank Bruno Antunes, Diana Santos, Nuno Cardoso and Cristina Mota for their valuable comments and suggestions.