A Rule-Based Approach for Arabic Named Entity Metonymy Resolution

Sadek Mansouri[1], Chahira Lhioui[1], Mounir Zrigui[1], Mbarek Charhad[2]

[1] LATICE Laboratory,
Research Department of Computer Science,
Tunisia

[2] Taibah University,
Saudi Arabia

mansouri sadek@hotmail.fr, chahira_m1983@yahoo.fr,
mounir.zrigui@fsm.rnu.tn, mbarek.charhad@gmail.com

Abstract. Named entity metonymy resolution aims to determine the correct meaning and type of named entity in a given context. This task has recently been subject to a growing interest for the Natural Language Processing community and mainly for Arabic language. In this paper, we have developed a new method to solve arabic named entity metonymy in text news. Our main contribution consists to propose a robust parser that extracts metonymy patterns of each entity using a set of contextual rules based on part-of-speech tags and semantic triggers. The performance of our proposed approach has been evaluated on a relatively large size corpus. The obtained results are promising and improve the named entity recognition system.

Keywords: Metonymy resoution, contextual information, arabic NE recognition.

1 Introduction

Everyday, the mass of stored information is becoming colossal and continuing to grow exponentially. Information is conveyed through different modes of communication such as video, audio and text. In this heterogeneous environment, the challenge is to facilitate user's fast and relevant access to the desired information, without being confused by the huge amount of data that is offered to him. Many Automatic Natural Language Processing (NLP) applications interested in developing methods and tools to respond to this challenge, such as the extraction of information, the search for information, indexing and automatic translation.

In these different domains, the task of named Entities Recognition(NER) plays a transversal role. The concept of named entity appeared in the mid-1990s as being a subtask of the information extraction activity. It consists of identifying certain textual objects such as person name, organization name and location. Over the years, research on these linguistic units focused on increasingly complex issues like disambiguation and enriched annotation but also on their recognition in different contexts.

Despite the performance of proposed approaches, the recognition of named entities is still a difficult task especially for the Arabic language. This is due to the specific features of Arabic text [1–3]:

– Lack of capitalization: Unlike Indo-European languages, Arabic language does not have the concept of capitalization. This represents a major obstacle for the Arabic language during the extraction of NE. In fact, capital latter is very effective in the proper names recognition process for certain languages as English or French. So its absence in the Arabic language imposes an urgent necessity to find alternatives and, ultimately, to use other conventional means such as lexicons, triggers words and grammatical rules.
– Complicated morphology: Arabic is a highly-inflected language. It uses an agglutinative strategy to form a word. If NE appears with agglutinative form, then this poses a difficulty for the identification of this entity.
– The absence of vowels: A non-vowelized word has many ambiguities in meaning or syntactic function. For example, the word can be a verb (go) to a voyellation or a proper name (Gold) for another. Thus, this example illustrates the impact of the vowels lack in words recognition.

In this work , we are particularly interested to solve the problem of metonymy of Arabic named entities. As a definition metonymy is a linguistic operation that allows the use of one entity to represent another which, as a result, leads to the emergence of polysemy phenomena for lexical units. Metonymy corresponds more exactly to a case where one entity seems to refers to many different NEs types which lead to an semantic ambiguity as shown in the following examples.

<div dir="rtl">تصفيات مونديال 2018: فرنسا تتصدر المجموعة الأولى</div>

Example 1. "FIFA World Cup Qualifiers 2018: **France** lead the first group".

<div dir="rtl">النادي الثقافي محمود المسعدي بتازركة</div>

Example 2. "Cultural Club **Mahmoud Masadi** Btazarka".

In the first example *France* refers to sport club not a place or country while in the second sentence *Mahmoud Masadi* refers to name of cultural club not to name of person. This type of ambiguity makes identification of NEs more difficult and raises NE disambiguation problem as one of the main challenges to research not only in the semantic web but also in areas of natural language processing in general. In this paper, we adress these challenges to solve the problem of metonymy for arabic named entity in text news.

To reach our objective, we define a several regular metonymic patterns for each entity class. We also have to establish a set of syntactic rules which provide to detect these patterns in given text and implemented with the linguistic platform Farassa. The rest of this paper is organized as follows: Section 2 provides insights from literature on related work. Section 3 presents the proposed approach. Section 4 discusses the approach evaluation and results, and Section 5 concludes the paper.

2   Related Works

The first work developed in the metonymy resolution (MR) task comes from the evaluation company SemEval 2007 [1] [4] and after by [5]. In this last work, the set of descriptors has been updated to include: the grammatical role of the potential metonymic word (PMW) (as subj, obj), determinant of PMW, grammatical number of PMW (singular, plural), number of words in PMW and the number of grammatical roles of PMW in a given context. The system proposed by [6] achieved the best results in terms of accuracy 85.2% using these features and a maximum entropic classifier.

In the same context, [7] reaches an accuracy of 85.1% using local characteristics of syntactic and global distribution generated with a suitable Xerox and proprietary deep analyzer. This system is based on an unsupervised approach with the use of contextual syntactic similarity calculated on large corpora such as the British National Corpus (BNC) with 100M tokens.

[8] use SVM (Support Vector Machine) with new descriptors (in addition to the descriptors provided by[5]) integrating grammatical collocations extracted from the BNC to learn selection preferences. Moreover, their system is based on external linguistic resources like WordNet 3.0, the Wikipedia category for optimization of the global context. In SemEval 2007, this system achieves an accuracy of 86.1%. Recently, [9] use a minimalist neural approach combined with a novel predicate window method for metonymy resolution.

Additionally, their system contribute with a new Wikipedia-based MR Dataset called RelocaR, which is tailored towards locations as well as improving previous deficiencies in annotation guidelines. The proposed system achieve 84.8% for the SemEval dataset. We remark that all proposed systems are designed to French and English language and to the best of our knowledge, our work present the first attempt to solve the problem of metonymy of arabic named entity.

3   Proposed Method

Figure 1 shows the global architecture of our proposed system. This system include three stages: Morpho-syntactic analysis, NE recognition step which integrates a divers set of lexical resources such as gazetteers and syntactic grammar. The final stage focuses on metonymy resolution to assign the correct type of an ambiguous named entity and enhance the initial results of NE recognition.

3.1   Morpho-Syntactic Analysis

In this phase, we segment input text into words based on spaces delimiter. Then we proceed by a morphological analysis of the corpus to extract useful information that will be used in the named entity recognition system. Given the agglutinate structure that has the majority of Arabic words, our morphological analyzer can separate and identify morpheme and associate all information necessary to the current treatment.

---

[1] http://nlp.cs.swarthmore.edu/semeval/tasks/index.php
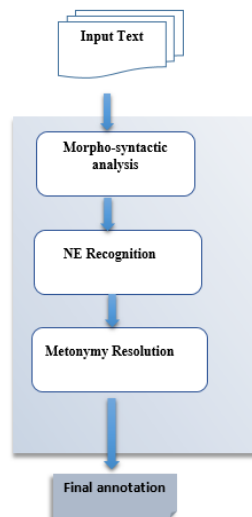
Fig. 1. The proposed system.

These forms are decomposed to recognize affixes (conjunctions, prepositions, pronouns, etc.) attached to it. These morphological possibilities in this analyzer can facilitate the identification of trigger words. As a matter of fact, each of these forms is associated with a set of useful linguistic information for the following step: lemma, grammatical labels, gender and number, syntactic information (such as: + Transitive) distributional information (such as: + Human), etc.

3.2 Named Entities Recognition

In the second step, we parse text files to detect named entities using Farassa tool[2]. Farasa is a fast and accurate text processing toolkit for Arabic named entities based on the CRF++[3] implementation of CRF sequence labeling. To improve NER in our datatset. We built a large gazetteer from Wikipedia. The gazetteer had 80,908 locations, 36,391 organizations, and 91,880 persons and 2000 event.

3.3 Metonymy Resolution

The task of metonymy resolution implies identifying the correct interpretation of a named entity in a given context. Some entities can be hard to annotate because of ambiguity between main types, such as locations, GPEs and organizations. Such entities can often take on different roles, according to their usage. The ACE guidelines describe two forms of metonymy.

---

[2] http://qatsdemo.cloudapp.net/farasa/demo.html

[3] http://code.google.com/p/crfpp/

Nickname metonymy occurs when the name of one entity is used to refer to another entity, such as a capital city referring to a government, or a location name denoting asports team. Cross-type metonymy occurs when multiple aspects of an entity are referenced at the same time, such as organizations and the facilities they occupy (e.g. They will be visiting the White House tomorrow). In this work, we intressed to metnoymy resolution. To this end, we first describe a serval regular metonymic patterns for each entity class and then we propose a syntactic grammars aiming to detect these patterns in arabic text news and solve the metnoymy problem.

Metonymic Patterns. In this section , we present a set of metonymic patterns which used to solve named entities ambiguity .

– Metonymic patterns of Organisation class
  1. **Organisation-to-event**:
  This Annotation is to be used when the name of the organization refers to an event organized by this organization.

  <div dir="rtl">المؤتمر الدولي لمنظمة الأمم المتحدة</div>

  "International Conference of **the United Nations**".

  2. **Organisation-to-Person**:
  This annotation is to be used when an organization name refers to people associated with it.

  <div dir="rtl">الناطق الرسمي لمنظمة الدفاع عن حقوق المستهلك</div>

  "The official spokesman of **the Organization for the Defense of Consumer Rights**".

  3. **Organisation-to-product**:
  Another widely used metonymy is Organisation-to-product, where the name of a commercial organization refers to its products.

  <div dir="rtl">سيارات فورد الشعبية تقتحم السوق العالمية</div>

  "**Ford's** popular cars are breaking into the global market".

  4. **Organisation-to-location**:
  This annotation is to be used when a company name referred to its own places.

  <div dir="rtl">أنفاق المقاومة الفلسطينية في قطاع غزة</div>

  "**Palestinian resistance tunnels** in the Gaza Strip".

– Metonymic patterns of location class similarly to organization names, place names can also refer not only to their primary reference, but also to other references related to it. The most frequent type is :
  1. **location-to-event**:
     We use this annotation when the country name refers to an event having taken place in this country.

الدورة العاشرة لأيام **قرطاج** السينمائية

"10th **Carthage** Film Festival".

2. **location-to-Organization**:
This annotation is to be used when a country name refers to people or organizations associated with it. We can think of when it comes to the government, a sports team or of the country's population.

تونس تترشح لمونديال 2018

"**Tunisia** to run for 2018 World Cup".

مجلس **النواب** يصادق على قانون المالية

"**The House of Representatives** approves the Finance Act".

– Metonymic patterns of Person class
1. **Person-to-time**
This annotation is to be used when a person's name refers to a period of time.

مصطفي الفيلالي وزير زمن **بورقيبة**

"Mustafa Filali Minister of Time of **Bourguiba**".

2. **Person-to-location**
This annotation is to be used when a person name referred to location.

المدرسة الإعدادية **علي بلهوان**

"Preparatory School **Ali Belhwan**".

3. **Person-to-Organization**
We apply this pattern when a person name referred to Organization.

جمعية **رحمة** لرعاية الأيتام بمنوبة

"**Rahma** Association for Orphans Care in Manouba".

Elaboration of Contextual Rules: Our focus in this part of study is to identify the contextual rules for each pattern. To this end, we have started to carry out a study of the training corpus. For all types of named entities, our focus was on the one hand, the types of grammatical relationships in which the units were involved. on the other hand, the lexical or contextual information attached to the arguments of these relations.

For each metonymy pattern, we therefore analyzed all of its occurrences (attached to a lexical unit) in the training corpus to derive grammatical and lexical configurations playing the role of metonymy resolution. This study is completed by the elaboration of contextual rules reflecting the discriminant configurations for each pattern. For example, let's take the following rule:"if the country name is followed by an verb so the pattron location-to-organization should be applied". The second contextual rule of location pattern is: "if the country name proceed by trigger reltad to event such as (Symposium, Forum, Conference), pattern of organization-to-event should be applied".

verb+   LOC

تونس نترشح لمونديال 2018

NE Recognition

ORG

تونس

Final annotation

Fig. 2. Example for location-to-organization rule.

LOC+ event trigger

مهرجان توزر الدولي

NE Recognition

Event

مهرجان توزر

Final annotation

Fig. 3. Example for location-to-event rule.

ORG+perp+ adj+event trigger

المؤتمر الدولي لمنظمة حقوق الإنسان

NE Recognition

Event

المؤتمر الدولي لمنظمة حقوق الإنسان
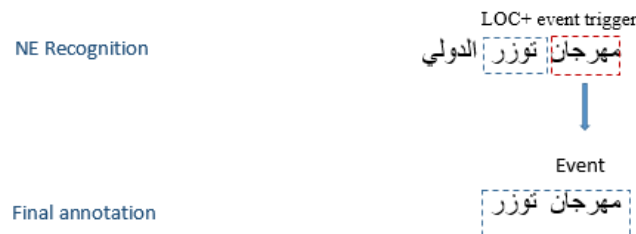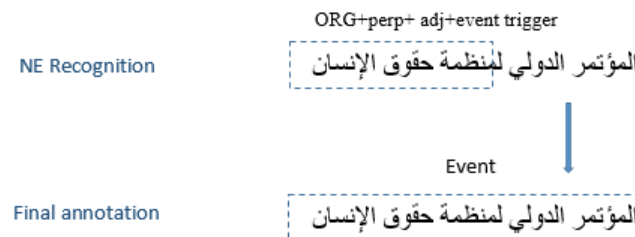
Final annotation

Fig. 4. Example for Organization-to-event rule.

For the Organization class, the most discriminating indices are prepositional phrases such as International Conference of the United Nations .To solve the problem , we apply this contextual rule :"if the country name proceed by trigger related to event such as (Symposium, Forum, Conference) + adj + preposition, pattern of organization-to-event should be applied".

For person class, the main problem of metonymy is caused by the fusion of meaning with other named entities. To solve this, we define this contextual rule: "if the name of person proceed by trigger related to location, so, we apply Person to location pattern."
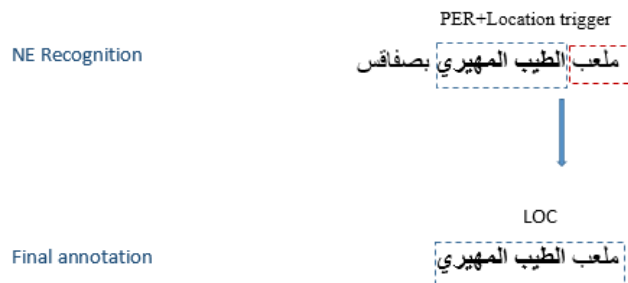
Fig. 5. Example for Person-to-location rule.

Table 1. Distribution of metonymy patterns.

Entity	Metonymy Petterns	%
Organization	Organization-to-event	7%
	Organization to Person	3%
	Organization-to-Location	5%
Location	location-to-event	3%
	location-to-organisation	12.3 %
Person	Person-to-location	5.6%
	Person-to-organization	12.3%
Total		37.6%

4 Evaluation and Results

4.1 Corpus

To the best of our knowledge, there is no reference benchmark dataset for Arabic named entities disambiguation However, to evaluate our approach we have assembled a large dataset on news articles extracted form Arabic Wikinews covering different domains of technology, politics and sport. Table 1 show the distribution of metonymic patterns in the training corpus and gives in idea about the complexity of the metonomy resolution task .It presents the number of metonymy cases for each named entity. In this table, we find that the total cases is equal to 37.6%.

4.2 Results

The evaluation of our system is as follows:

– The texts of evaluation corpus are all manually annotated.
– After we have automatically annotated these texts using our named entities recognition system.
– Then we have applied our method of metonymy resolution to enhance the initial results.

Table 2. Annotation results.

Entity	without (MR)	with (MR)
Organization	78,90%	85,5%
Person	81,25%	82,5%
Location	81,91%	83%

In the final step ,we have established a comparison between these annotations with corpus quality assurance. The overall evaluation of entities extracted by our system is based on the use of F-measure. This measure combines the precision and the recall. Precision measurement is defined by the percentage of entities found by the system and which are correct. The recall is defined as the ratio between the numbers of found correct entities by the number of entities extracted from the reference articles.

Table 2 shows the obtained results for different entities with and without metonymy resolution (MR).We remark that our proposed method(MR) improve the f measure rate of the system by 10.05 % particularly organization entity achieved the best result.This due to the strong presence of metonymy patterns of Person to Organization and Person to Organization in dataset and the performance of our proposed syntactic grammar.

5   Conclusion

The treatment of lexical metonymy particularly the metonymy of named entities, can improve a number of treatments, among which the extraction tasks information and questions and answers. Our work present the first attempt to solve the problem of metonymy for Arabic named entities. The main contribution is to define the metonymy patterns and syntactic grammar aiming to find the exact type of each entity according to the context where is used. The experimentation and evaluation results are promising. In future works, we can test our system on a larger corpus and try to define new metonymy patterns to improve our proposed approach.

References

1. Mansouri, S., Lhioui C., Charhad, M., Zrigui, M.: Text-to-Concept: A Semantic Indexing Framework for Arabic News Videos. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2017 (2018)
2. Lhioui, C., Zouaghi, A., Zrigui, M.: A Rule-based Semantic Frame Annotation of Arabic Speech Turns for Automatic Dialogue Analysis, Procedia Computer Science, vol. 117, pp. 46–54 (2017)
3. Lhioui, C., Zouaghi, A., Zrigui, M.: Realization of Minimum Discursive Units Segmentation of Arab Oral Utterances. International Journal of. Computational Linguistics and Applications. vol. 7, no. 1, pp. 31–50 (2016)
4. Markert, K., Nissim, M.: SemEval-2007 Task 08: Metonymy Resolution at SemEval-2007, pp. 36–41 (2007)
5. Markert, K., Nissim, M.: Data and models for metonymy resolution. Lang. Resour. Evaluation, vol. 43, no. 2, pp 123–138 (2009)

6. Farkas, R., Simon E., Szarvas, G., Varga, D.: GYDER: Maxent Metonymy Resolution. pp. 161–164 (2007)

7. Brun. C., Ehrmann, M., Jacquet, G.: A Hybrid System for Named Entity Metonymy Resolution. LTC 2007: pp. 118–130 (2007)

8. Nastase, V., Strube, M.: Combining Collocations, Lexical and Encyclopedic Knowledge for Metonymy Resolution. EMNLP 2009: pp. 910–918 (2007)

9. Gritta, M., Taher-Pilehvar, M., Limsopatham, N., Collier, N.: Vancouver Welcomes You! Minimalist Location Metonymy Resolution. vol. ACL, no. 1, pp. 1248–1259 (2017)

10. Mahmoud, A., Zrigui, M.: Semantic Similarity Analysis for Paraphrase Identification in Arabic Texts. PACLIC 2017, pp. 274–281 (2017)

11. Zouaghi, A., Zrigui, M., Antoniadis, G.:Compréhension automatique de la parole arabe spontanée. Traitement Automatique des Langues, (2008)

12. Ayadi, R., Maraoui, M., Zrigui, M.: Latent topic model for indexing arabic documents. International Journal of Information Retrieval Research (IJIRR), vol. 4, no. 2, pp. 57–72 (2014)

13. Merhben, L., Zouaghi, A., Zrigui, M.: Lexical Disambiguation of Arabic Language: An Experimental Study Polibits, pp. 49–54 (2012)

14. Mohamed, M., Mallat, S., Nahdi, M. A., Zrigui, M.: Exploring the potential of schemes in building NLP tools for Arabic language. International Arab Journal of Information Technology (IAJIT), vol. 12, no. 6 (2015)

15. Mansouri, S., Charhad, M., Zrigui, M.: Arabic Text Detection in News Video Based on Line Segment Detector. Research in Computing Science. vol. 132, pp. 97–106 (2017).

16. Slimi, J., Mansouri, S., Ben-Ammar, A., Alimi, A. M.: Semantic browsing in large scale videos collection. OAIR 2013, pp. 53–56 (2013).

17. Slimi, J., Mansouri, S., Ben-Ammar, A., Alimi, A. M.: Video exploration tool based on semantic network. OAIR 2013, pp. 213–214