Towards a State-of-The-Art Report on the Hybrid Processing of MWEs

Mike Rosner[*], mike.rosner@um.edu.mt
Matthieu Constant[†], Matthieu.Constant@u-pem.fr

PARSEME WG3:
April 2014

Abstract

This report began as a summary report of the WG3 sessions that took place at the Athens meeting in March but grew into something slightly more ambitious. It describes some modifications that were made to the original aims and objectives and a classification scheme for work being carried out within the group. It then sets out a plan of action for the second year of the project. This includes the production of a state-of-the-art report and some suggestions on how the group might contribute to future meetings in Haifa and Malta.

1 Introduction

WG3, whose membership has now reached about 25[1] sets out to investigate the use of hybrid methods to increase the efficiency and accuracy of parsing MWEs. There is a lively ongoing discussion on the terminology used to describe the ingredients of hybrid methods. Some refer to these as "data driven" versus "rule based", others to "probabilistic" versus "linguistic". A recent contribution to this debate casts the distinction according to whether the constraints that contribute to a model are *type* or *token* constraints [3].

Despite the discussion, there is general agreement that such methods involve various combinations of two fundamentally different approaches, namely (i) "cognitively motivated theories of language in the tradition of generative linguistics, with introspection as primary evidence" and (ii) "approaches motivated by empirical coverage, with collections of naturally occurring data as primary evidence" [2]. We shall refer to these as symbolic and statistical approaches respectively.

[*]University of Malta
[†]Université de Marne-la-Vallée
[1]as at March 2014

During the first general meeting in Warsaw, there was some discussion on the scope of WG3. In particular concerning (i) whether statistical parsing of MWEs should fall within WG3 or WG2 and (ii) whether to include Machine Translation of MWEs in WG3, Hybrid MT being a major theme of current research (see [1]) and MWEs being a major concern of MT.

1.1 Revising WG3 Objectives and Outcomes

To resolve these issues the scope of WG3 was slightly modified as reflected in its present title: "Statistical, Hybrid and Multilingual Processing of MWEs". The objectives now include

- Elaborating the notion of hybridity.
- Improving our understanding of how these may be applied to the processing of MWEs.
- Investigating the relation between hybrid processing methods and multilingual applications.

Apart from the multi-lingual element the intended outcomes remain substantially as they appear in the MOU, namely recommendations of best practices for

- enhancing statistical parsing with linguistically motivated resources such as MWE lexicons and valence dictionaries, e.g. by MWE-oriented reranking of state-of-the-art parsers results;
- enhancing symbolic parsing of MWEs with probabilistic scores, in order to avoid spurious syntactic ambiguities while parsing MWEs;
- guidelines for the extraction of statistical information from various unlabelled data sources, parallel corpora, automatically annotated corpora, treebanks and for encoding it in lexicons.

This report is mostly about the first of these objectives, and in order to shed light on the nature of hybridity, we decided to look in particular at the poster submissions to this working group for the second meeting at Athens.

2 A Classification Scheme for Work under WG3

Below we propose a simple classification scheme for research being carried out in WG3, and given the broadened definition of WG3, this is not an entirely trivial task. For this reason we decided to start by looking at something concrete, namely the posters that were submitted to this group.

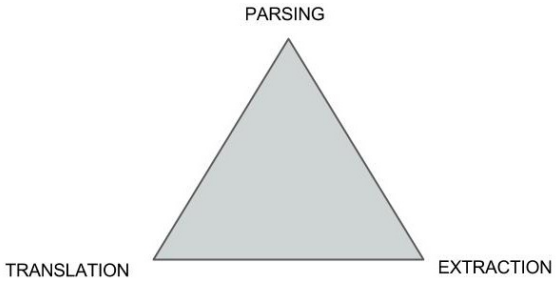

Figure 1: Classifying Poster Presentations

2.1 Submitted Posters

The poster presentations submitted to WG3 clearly concern a number of different processing issues. A cursory glance at the poster titles (see Appendix A) suggests a strong correlation with three broad themes shown in figure 2.1 according to the appearance of certain keywords.

- **Parsing** is mainly concerned with the definition of *parsing algorithms* that produce output in which MWEs are explicitly marked as such. A key characteristic of all such algorithms is that they must in some way *incorporate MWE resources* whose existence is presupposed. The fact that there are many ways to represent such resources and to incorporate them gives rise to wide variation in the design of such algorithms.

- In contrast to this, the main goal of **Extraction** is the *creation* MWE resources by some kind of discovery process which operates on unlabelled or partially labelled data. This can be achieved by adding the relevant annotations and/or by extracting identified MWEs into a lexicon.

- **Translation** is something of a mixture, involving both the incorporation and creation of MWE resources. This is because translation, rather like parsing, involves an algorithmic component which is reflected in the design of algorithms which explicitly incorporate MWE resources into the translation process. The complication here is that the resources may be bilingual. Algorithms are also involved in the creation of such resources. An example would be the discovery of bilingual phrasal pairs containing at least one MWE like `kick/donner un coup de pied` from aligned but unlabelled data using statistical methods.

2.2 An Initial Classification Scheme for Hybrid Processing

This section suggests a slightly more refined scheme for classifying the work being carried out under the aegis of WG3 that is motivated by the considerations

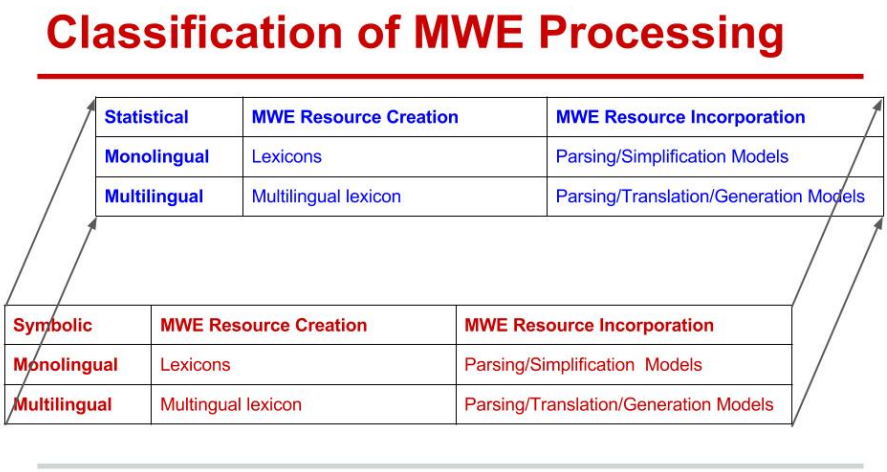

Figure 2: Classification of Hybrid Processing

of section 2.1. The proposed scheme hinges on three dimensions into which the three themes of extraction, parsing and translation can be fitted.

1. Resource Creation/Resource Incorporation
2. Monolingual-Multilingual
3. Symbolic-Statistical

These are depicted in Figure 2 on the x, y and z axes respectively. The x-axis represents the role of MWE resources, the y-axis the degree of multilinguality. The outer boxes are the names of the axes, whilst the inner boxes mention data resources that are the output of creation, and processes that incorporate such resources.

A key addition is the z-axis which depicts the methodological framework. It is here that we incorporate the notion of hybridity. In the figure, the far end of the axis (blue) represents purely statistical approaches and purely symbolic approaches respectively. The challenge is how to label the intermediate points on the scale.

This scheme is not intended to be the final word, but rather a useful initial guideline to facilitate collaborative work within the group. For example, if two individuals are concerned with the creation of monolingual MWE resources for their respective languages, then it is likely that their respective work will share some common themes - for example concerning representation or methods of MWE identification.

The next section discusses our proposal for the compilation of a state-of-the-art report during the current year of the action. We regard this as a logical

next-step based on the proposed classification scheme to help members of the group discover the people with whom they can fruitfully exchange ideas.

3 An Action Plan for Year 2

In this section we discuss a possible plan for the the second year of the programme which will be structured around the following items:

- State-of-the-art report
- WG3 at Haifa
- Malta Meeting (March 2015)

Now we discuss each of these in turn.

3.1 State-of-the-Art Report (SOAR)

As mentioned above, a SOAR is an important tool to help us collaborate on specific topics within the group. Before this happens, it is crucial that the group collaborate in a more general sense and as a whole on the compilation of the SOAR. A substantial part of the content will be short summaries of ongoing work - for example as already reported in the Posters - together with their classification in the scheme suggested above which if necessary can be adjusted.

Proposed Report Structure

Here is an initial proposal for the structure of the report:

I Introduction

II Suggested Classification Scheme

III Individual Contributions

IV Contrastive Analysis of the Contributions[2]

V Summary of SOA

VI Expected Future Work

This report will serve two purposes. The first is to contribute to the annual report of the project as a whole. The second is to help provide structure for our contribution to the Haifa meeting. A hoped-for outcome is the emergence of subgroups within WG3 who might be in a position to collaborate by co-authoring papers, producing common resources etc. and these might be made the focus of one or more sessions at Haifa.

[2]The contrastive analysis could be, roughly, a large two-dimensional table with contributions indexing lines and aspects indexing columns.

3.2 Outline for Haifa Meeting

It is as yet rather early to determine the way in which the Haifa meeting will proceed, but we assume that this will involve at least one WG3 session involving some joint work from WG3 participants.

3.3 Timetable

End-April:

- Publication of this document as an internal report.
- Call for short (half-page) submissions from all WG members including current poster authors. Each contribution will position itself with respect to an updated version of classification scheme that will be manifest as an online form whose results will be collected.

End-May:

- Assessment of contributions and adjustment of classification scheme as appropriate.
- Draft version of SOAR
- Identification of topics for WG3 subgroups

End-July:

- Final (hopefully publishable) version of SOAR
- Draft documents from groups

4 Acknowledgement

The authors thank Agata Savary for comments on the first draft of this paper.

References

[1] M. Costa-Jussá, R. Banchs, R. Rapp, P. Lambert, K. Eberle, and B. Babych. Workshop on hybrid approaches to translation, overview and developments. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 1–6. ACL, August, 2013. `http://aclweb.org/anthology//W/W13/W13-2801.pdf`.

[2] Judith L. Klavans and Philip Resnik, editors. *The Balancing Act: combining symbolic and statistical approaches to language*. MIT Press, Cambridge, MA, 1996.

[3] Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan T. McDonald, and Joakim Nivre. Token and type constraints for cross-lingual part-of-speech tagging. *TACL*, 1:1–12, 2013.

5 Appendix A: Titles of Posters submitted to WG3 at Athens Meeting

PARSING

Dimitrios Kokkinakis "Swedish multiword expressions and sublanguage parsing" (WG 2, 1, 3, 4)

Gerold Schneider "Improving PP attachment in a hybrid dependency parser using semantic, distributional and lexical resources" (WG 3)

Joakim Nivre "Transition-Based Parsing with Multiword Expressions" (WG 3)

Istvn Nagy T., Veronika Vincze "Detecting Multiword Expressions by Dependency Parsing" (WG 3)

EXTRACTION

Markus Egg, Will Roberts, Valia Kordoni "Multiword Expression Identification for German" (WG 1, 3)

Amalia Todirascu "A Hybrid Multilingual Method to Extract Collocations from Corpora" (WG 3)

Yaakov HaCohen-Kerner "A ML research proposal for detecting Multi-Word Expressions" (WG 3)

Federico Sangati, Andreas van Cranenburgh "Identifying Multi-Word Expressions in Large Treebanks with Tree Kernels" (WG 3, 4)

Carla Parra Escartn, Hctor Martnez Alonso "Compound dictionary extraction and WordNet. A dangerous liaison?" (WG 3)

TRANSLATION

Johanna Monti "A knowledge-based approach to multiwords processing in machine translation: the English-Italian dictionary of multiwords" (WG 1, 3)

Carla Parra Escartn, Stephan Peitz, Hermann Ney "Linguistics, German Compounds and Statistical Machine Translation. Can they all get along?" (WG 3)

(OUTLIER)

Martin Emms, Arun Jayapal "Sense changes and Multiword Expressions" (WG 3)