PARSEME WG3

Statistical, Hybrid and Multilingual Processing of MWEs

Athens, 11th March 2014

Mike Rosner, University of Malta
Matthieu Constant, Université de Marne-la-Vallée

Working Group Session: Discussion Points

Outline

Session 1: 11.00-12.30

- Introduction: About WG3
- Presentations
    - Joakim Nivre, *Transition-Based Parsing with Multiword Expressions*
    - Ramona Enache, *Building Bilingual Multiword Lexicons from Parallel Text*
- Discussion: how does hybrid processing enter into the above systems?

Session 2: 14.00-15.30

- Classification of WG3-related poster proposals
- WG3 outcomes
- Roadmap
- Actions

About WG3

Currently 25 Members

Original name:

- Hybrid Parsing of MWEs

Hybrid = statistical + rule-based

Problems noted in Warsaw meeting

- Current work on MWEs not exclusively about parsing
- Multilingual dimension not included
- Confusion (e.g. wrt statistical processing) with role of WG2

WG3 revised

Statistical, Hybrid and Multilingual Processing of MWEs

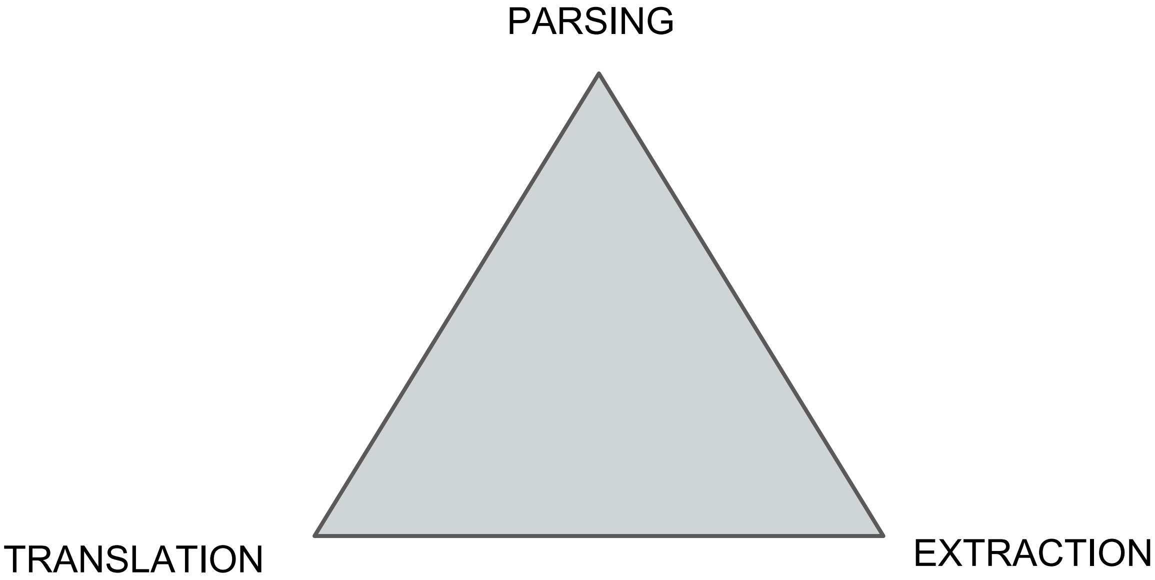
with additional objectives

- Elaborate ways of combining data-driven and knowledge-based methods to yield various hybrid processing models.
- Improve our understanding of how these may be applied to the processing of MWEs.
- Investigate relation between hybrid processing methods and multilingual applications.

Issues

- Nature of hybridity
- Processing: includes parsing but also other things
- Role of multilinguality

Processing MWEs: 3 Main Themes



PARSING

TRANSLATION

EXTRACTION

WG3-related posters: themes

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)

*István 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 Escartín, Héctor Martínez 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 Escartín, Stephan Peitz, Hermann Ney "Linguistics, German Compounds and Statistical Machine **Translation**. Can they all get along?" (WG 3)

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

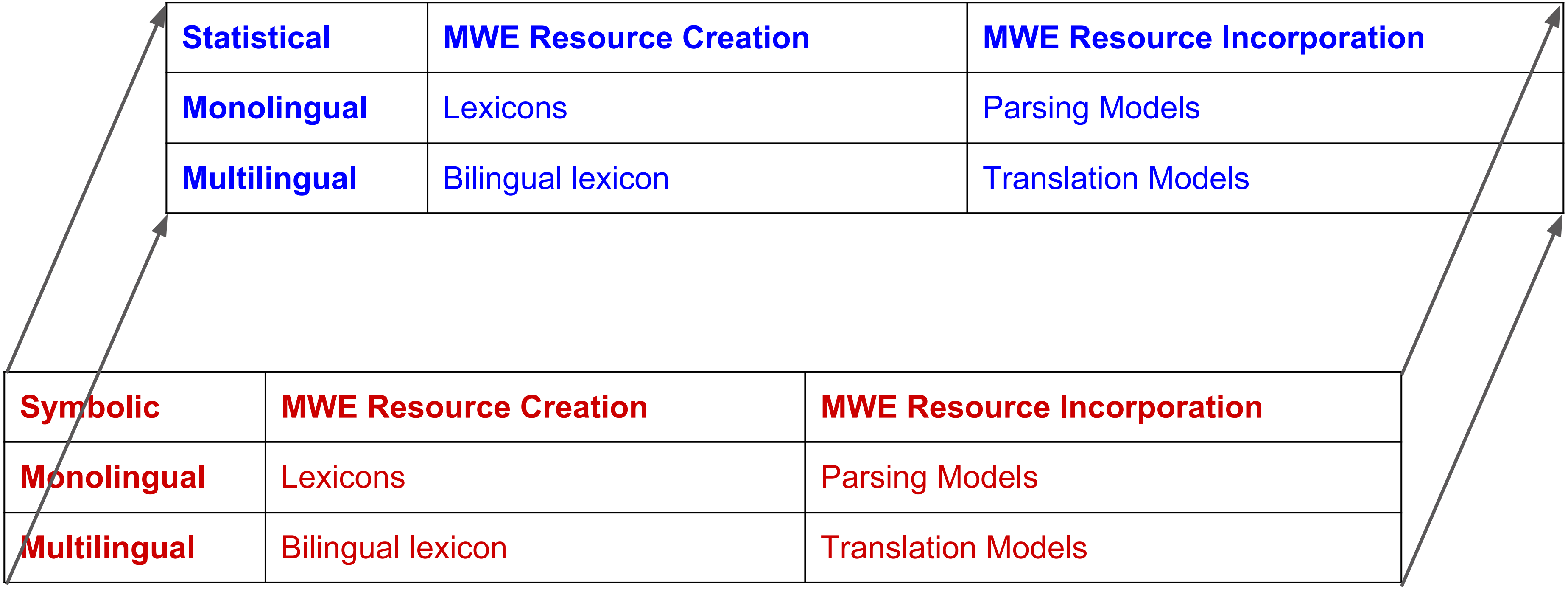
Classification of MWE Processing

THREE MAIN AXES FOR CLASSIFYING WORK

1. Monolingual/Multilingual
2. Resource Creation/Resource Incorporation
3. Hybridity: Symbolic/Data-Driven

	MWE Resource Creation	**MWE Resource Incorporation**
Monolingual	Lexicons	Parsing Models
Multilingual	Bilingual lexicon	Translation Models

Classification of MWE Processing



Statistical	MWE Resource Creation	MWE Resource Incorporation
Monolingual	Lexicons	Parsing Models
Multilingual	Bilingual lexicon	Translation Models

Symbolic	MWE Resource Creation	MWE Resource Incorporation
Monolingual	Lexicons	Parsing Models
Multilingual	Bilingual lexicon	Translation Models

Intended Outcomes (from MOU)

- Recommendations of best practices for enhancing data-driven parsing with linguistic resources such as MWE lexicons and valence dictionaries, e.g. by MWE-oriented reranking of state-of-the-art parsers’ results;
- recommendations of best practices for enhancing knowledge-based parsing of MWEs with probabilistic scores, in order to avoid spurious syntactic ambiguities while parsing MWEs,
- guidelines for extracting probabilistic scores from treebanks and for encoding them in lexicons (cf. WG1 and WG4).

Action Plan

- May 2014: State of the art report
- Suggested structure
- Year 2
  - Goals for year 2
  - WG3 at next general meeting

Suggested Report Structure

1. Introduction
2. Classification scheme for hybrid processing
3. Current SOA
    a. contributions by authors of WG3 posters within scheme. Standard format for contributions?
    b. identify gaps - other contributions
4. Conclusion, future work

Discussion 1

GENERAL

- Classification is useful because it helps to identify work by different people on the same problem.
- As a WG we might try to encourage joint work by collaborators in each category (subgroups)?
- Nevertheless there is still great diversity in the ways people work and it might be premature to expect them to work together (Wehrli)

CLASSIFICATION

- Rule-based/statistics-based is a false distinction. Symbolic versus data-driven approaches better. But in reality we are looking for a compromise between hard constraints (e.g. that output will be in the form of a dependency tree) and soft constraints (e.g. where we estimate the probability of a given output for a given sentence). (Nivre)
- Also consider distinction between working at token level v. type level? Annotation always works at token level whilst grammar rules express generalisations over types. See TACL 2013 paper about this distinction. (Nivre).
- In proposed classification table, multilingual aspect also involves multiword-aware parsing (audience) -> TODO: correct in classification table: translation models -> Parsing/translation models
- symbolic vs. statistical -> binary vs. numerical? (audience)
- Be aware: not throw away "distributional semantics" and compositionality issues (audience)

Discussion 2

OUTCOMES

extracting probabilistic scores from treebanks means extracting statistical information from treebanks/automatically annotated data/unlabeled data/parallel corpus/parallel treebanks?

STATE OF THE ART REPORT

- Classification scheme should work as a collaborative framework

Actions:

- Mid April: Mike circulates updated version of classification scheme
- Mid-May: Poster contributors should write a half-page contribution by mid-May. Each contribution
    - positions itself with respect to proposed classification scheme if possible
    - contains short summary of their contribution