Processing MWEs in Machine Translation Systems

Amalia Todirascu, University of Strasbourg, France, Johanna Monti, University of Sassari, Italy

WG3/WG 3.2 MWE and Translation

Multiword Expressions (MWEs)

- ● lexical items composed of several lexemes which display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity (Baldwin and Kim (2010) and (Sag *et al.*, 2002)
- ● several categories (idiomatic expressions, proverbs, collocations, compound words, domain-specific terms, named entities, lexical bundles, …)
- ● Fixedness vs. variability (morpho-syntactic, syntactic)
- ● Non-compositional vs compositional meaning

Translation problems (Hurskainen, 2008)

Word-for-word translation strategy fails:

□MWE non-compositionality

□Strong lexical preferences

□Morpho-syntactic and syntactic variability

Specific translation strategies for:

□ Compounds

□ Named entities

□ Terms

Statistical MT systems

Frequency criteria

□Phrase based tables => detecting fixed MWE

□ Hybrid methods

– linguistic and contextual properties of MWE

– frequency criteria and morpho-syntactic properties

□ Linguistic rules

– POS tags and lemmas (Bouamor *et al.*, 2012) (Caseli *et al.*, 2009)

– Syntactic information (Venkatapathy and Joshi, 2006)

Example based MT systems

•MWE translation examples as linguistic rules (Franz *et al.* 2000), (Gangadharaiah and Balakrishnan, 2006)

Rule based MT systems

□ lexical approaches to identify and translate contiguous MWE, using existing monolingual or bilingual dictionaries

□ compositional approaches to identify and translate MWE using linguistic rules OpenLogos (Scott, 2003; Scott and Barreiro, 2009; Barreiro *et al.*, 2011)

Integrating MWE in MT systems

□ Preprocessing MWE before word alignment

➔ Using external resources

➔ Specific tools

■ Linguistic methods (Wehrli *et al.*, 2009)(Seretan, 2009)

■ Statistical methods (Smadja, 1993, Evert, 2005)

■ Hybrid methods (Ramisch *et al.*, 2013)

□ MWE alignment

➢Methods using simple word alignment to find MWE alignment (Melamed, 1997) (Okita *et al.*, 2012)

➢Statistical measures to find parallel MWE (Pal *et al.*, 2013)

Specific strategies

□**Terms** (Bouamor *et al*, 2012):
Monolingual term extraction tools and alignment (Macken *et al.*, 2008), (Kontonatsios, 2014)
External resources (Wu *et al.*, 2008)

□ **Named Entities** (Tan and Pal, 2014):
NER recognizers
Multilingual collections

□**Compounds**:
splitting method and the compounds are added to the training corpus (Weller *et al.*, 2014) (Cap *et al.*, 2014), (Ullman and Nivre, 2014)
Rule-based methods

Evaluation

□ MWE annotation is difficult (variability, discontinous MWE)

□ Lack of annotated resources

□ Hand-made corpus representative for specific phenomena: verb compounds for English and French (Ramisch *et al.*, 2013) aligned, sometimes discontinuous verb+noun collocations for French and Romanian (Navlea, 2014), compound nouns for English and German (Parra Escartin *et al.*, 2014), Support verb constructions (Barreiro *et al.* 2014)

□ Comparison of MWE processing in different MT systems (Barreiro *et al.* 2013 and 2014, Monti *et al.* 2011 and 2013)