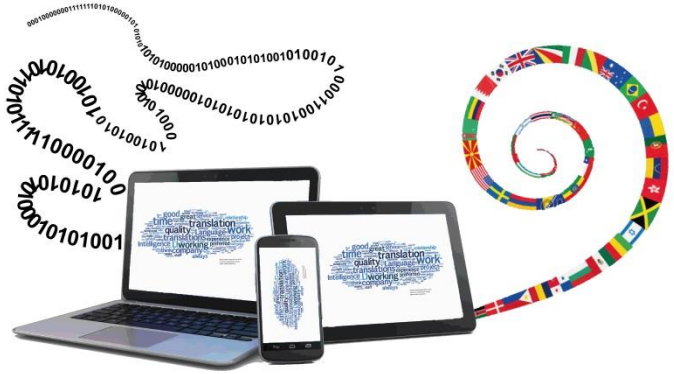
MWE processing in Machine Translation: State of the Art and Open Challenges

Johanna Monti
Department of Human and Social Sciences
University of Sassari
Italy

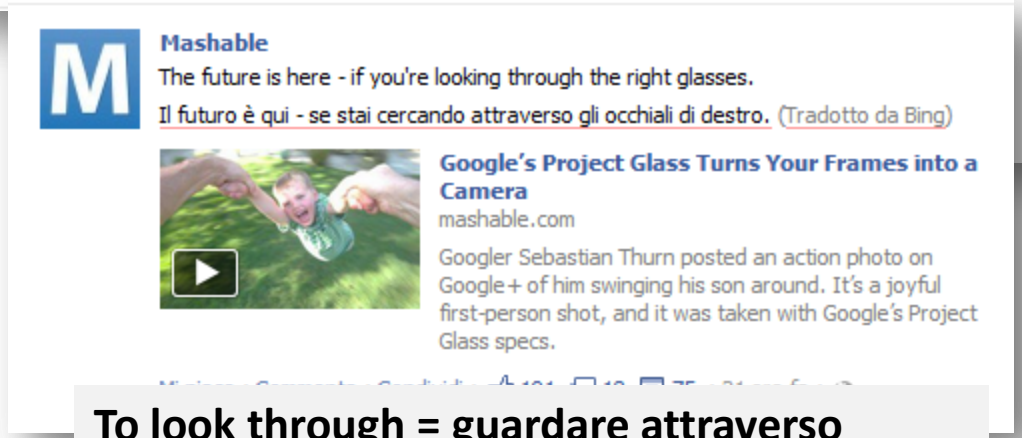


PARSEME meeting – Warsaw 16-19 September 2013

MWE: a hard nut to crack for MT?

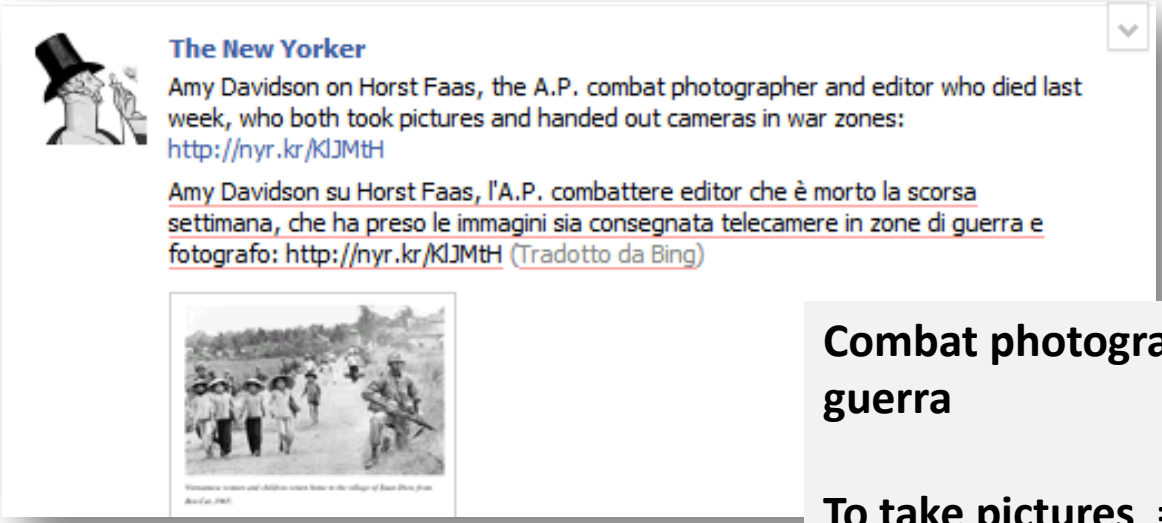
to go to the dogs= andare in malora

to get married = sposarsi



To look through = guardare attraverso

MWE: a hard nut to crack for MT?



Combat photographer = reporter di guerra

To take pictures = fotografare

MWE: a hard nut to crack for MT?

europeana
think culture

Search

Return to search results

© Rights reserved - Free access

View item at
Portable Antiquities

Share

Cite on Wikipedia

Translate details

Select language

Powered by Microsoft® Translator

AMPHORA

Description: Fragment of earthenware amphora base, with four concentric ridges running parallel from the base upwards, and fossilised marine organisms on the surface, indicating that it has been under water for some time. The fabric has been high fired so that it is hard, almost like stoneware, with traces of copper green glaze on the interior of the vessel, and inclusions which, along with its shape, suggest that it was used for olive
late 16th to mid 17th century. Hurst, Neal & Var
similar example on page 65, Fig.29, No.81.

Creator: Anna Tyacke

Contributor: CORN

Geographic coverage: KERRIER

Time period: 1575 1650

Type: Image

Format: text/html

Subject: archaeology; http://www.eionet.europa

Identifier: http://www.finds.org.uk/database/arte

Is part of: Portable Antiquities Scheme - Finds

Language: en-GB

Publisher: The Portable Antiquities Scheme

Data provider: Portable Antiquities

Provider: CultureGrid

Providing country: United Kingdom

© Rights reserved - Free access

View item at
Portable Antiquities

Share

Cite on Wikipedia

Translate details

Italian

Powered by Microsoft® Translator

Return to original language

AMPHORA

Description: Frammento di anfora di terracotta base, con quattro creste concentriche che corre parallelo dalla base verso l'alto e gli organismi marini fossilizzati sulla superficie, che indica che è stato sotto l'acqua per qualche tempo. Il tessuto è stato alto sparato così che è difficile, quasi come gres porcellanato, con tracce di smalto verde rame all'interno della nave e inclusioni che, insieme con la sua forma, suggeriscono che è stato usato per l'olio d'oliva e fatto a Siviglia nel tardo XVI alla metà del XVII secolo. Hurst, Neal

Creator: Anna Tyacke

Contributor: CORN

Geographic coverage: KERRIER

Time period: 1575 1650

Type: Immagine

Format: testo/html

Subject: Archeologia; http://www.Eionet.Europa.eu/GEMET/concept/530

Identifier: http://www.finds.org.uk/database/artefacts/record/id/112239

Is part of: Portable Antiquities Scheme - Finds

Language: en-GB

Publisher: The Portable Antiquities Scheme

Data provider: Portable Antiquities

Provider: CultureGrid

MWE in MT: still a problem!

Very frequent phenomenon both in standard and domain specific languages

Unpredictable translations →mistranslation in MT due to fragmentation and literal translations

Semantic idiosyncracy: different degrees of compositionality

Morpho-syntactic idiosyncracy: formal variations with dependencies of elements even if non-contiguous

Back to the past?

«The only way for a machine to treat idioms successfully is not to have idioms» (Bar-Hillel 1952)

MWE: different types

Idiomatic expressions

- En. *Hold your tongue* (→ It. *Frena la tua lingua*)

Proverbs

- En. *The early bird gets the worm* (→ It. *Chi dorme non piglia pesci*)

Word compounds

- adjectives = En. *good-looking* → It. *di bell'aspetto*
- prepositions = En. *in order to* → It. *per*
- adverbs = En. *arm in arm* → It. *sottobraccio*
- conjunctions = En. *in spite of* → It. *nonostante*.
- Nouns = En. *credit card* → It. *carta di credito*
- Verbs = En. *Kick the bucket* → It. *tirare le cuoia*

Noun compounds

- Nouns modified by:
  - nouns = En. *credit card* → It. *carta di credito,*
  - Adjectives = En. *perfect pitch* → It. *orecchio assoluto*
  - adjectival locutions = En. *amount of time* → It. *quantità di tempo*.

MWE: different types

Verb compounds

- Verbs modified by:
  - particles = En. *give up* → It. *rinunciare*,
  - prepositions = En. *adapt to* → It. *adattarsi a*,
  - nouns = En. *advance a project* → It. *presentare un progetto*
- Phrasal verbs
  - En. *give away* → It. *dar via, donare*;
  - En. *give back* → It. *restituire, rendere, ridare*;
  - En. *give in* → It. *consegnare, arrendersi;*
  - *En. give off → It. emettere, sprigionare;*
  - *En. give out → It. distribuire;*
  - En. *give over* → It. *dedicare, consegnare*;
  - En. *give up* → It. *cedere, arrendersi, smettere*;
  - En. *give way* → It. *cedere*
- Light verbs or support verb constructions
  - En. = *to give a presentation (to present)* → It. *Fare una presentazione (presentare)*

MWE: different types

Term compounds

- Various types of domain-specific compounds, but mainly noun compounds.
- close relationship between terminology and multi-words and, in particular, word compounds (90% of domain-specific terms)
- mono-referential : *pay scale* in the financial domain, = "the different levels of pay for a particular job, relating to different degrees of skill or experience" → *It. scala dei salari.*
- Their meaning, cannot be directly inferred by a non-expert from the different elements of the compounds.
- Complex and varied morpho-syntactic and semantic behaviour with unpredictable translations.

MWE: different types

Collocations

- Any statistically significant co-occurrence" of words (Sag et al. 2002) : non-casual, restricted, arbitrary and recognisable combinations of words (collocates)
- usually semantically compositional,
- particularly relevant in MT since they cannot always be translated literally: En. *anticipate the salary* → It. *anticipare lo stipendio*; En. *anticipate a pleasure* → It. *pregustare un piacere*; En. *anticipate Ving* → IT. *prevedere di Vinf*.
- Unpredictability of word co-occurrence on the basis of syntactic or semantic rules: En. *I did my homework* vs **I made my homework*.
- The translation of collocations requires a correct interpretation of their meaning which is determined by the co-text. En. *anticipate a pleasure* → *It * anticipare un piacere* (Google Translate).

Other types of MWE

- Named Entities: En. *Economic Council*
- Lexical bundles: En: *I believe that, as much if not more than, if I were you*

MWE as part of a continuum

Elia & D'Agostino (1998) identify 4 types of combinations of phrases or sentences:

high degree of variability of co-occurrence
(free internal distribution)
dirty water, clean water

limited degree of variability of co-occurrence
(restricted internal distribution)
natural water

3. **no or almost no variability of co-occurrence**
(fixed internal distribution)
heavy water

4. no variability of co-occurrence (proverbs)
all good things come to he who waits

Some properties

Non-substitutability: *in deep water → in hot water*; *gas chamber → *gas room*

Non-expandability: *get a head start → *get a quick head start*

Non- reducibility: *take advantage → *what did you take? advantage*; *Did you take it?;*

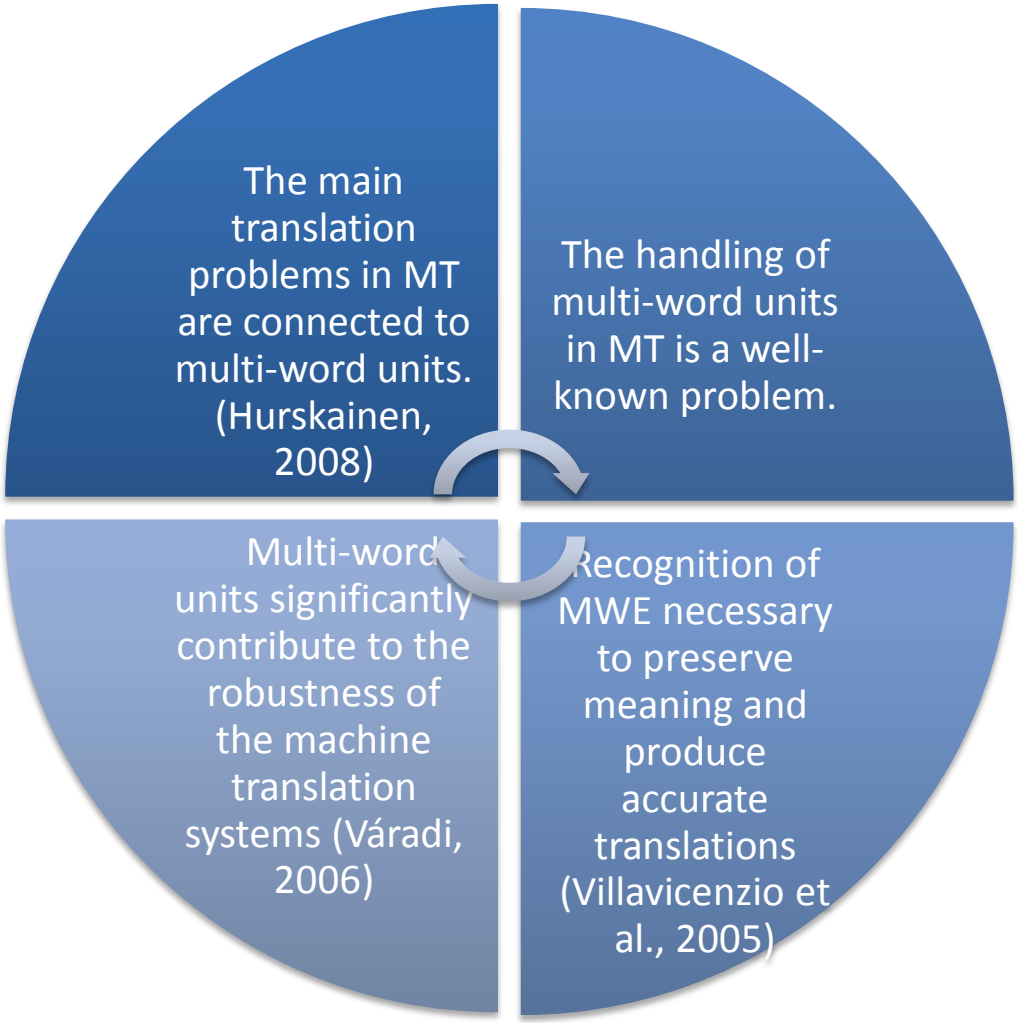
Invariability: *fish out of water → *fishes out of water*; *dead on arrival → *dead on arrivals*; *in high places → *in high place; credit card → *card of credit*;

Non- displaceability: *wild card → *is wild this card?- back and forth → *forth and back*

Institutionalisation of use: *in tempo reale* (a loan translation of the English expression *in real time*) vs **in tempo irreale* (**in unreal time*)

Non-translatability: E n. *It's raining cats and dogs* → It. **Sta piovendo cani e gatti*, It. *compilare un modulo* → En. **Compile a module*

MWE in Machine Translation



Different approaches to MWE

Rule-Based Machine Translation (RBMT)

- **Lexical approach** (electronic dictionaries) → suitable for contiguous MWE
- **Compositional approach** (rules) → useful for translating MWE not coded in the system dictionary and for translating verbal constructions

Example-based Machine Translation (EBMT)

- **Analogy principle** – reuse of translation stored in the system
- **Alignment** - uses raw (un-annotated) input data to extract correspondences from large parallel corpora
- **Sub-sentential alignment** from parallel bilingual corpora

Johanna Monti MWE processing in Machine Translation: State of the Art and Open Challenges
PARSEME meeting – Warsaw 16-19 September 2013

Different approaches to MWE

Statistical Machine Translation (SMT)

- MWE as a problem of **automatically learning** and **integrating translations** of very specific MWE categories, **word alignment**, or **word sense disambiguation (WSD)**
- Able to identify MWE with no or almost no variability in co-occurrence among words
- Shortcomings in identifying MWE with a high and limited degree of variability of co-occurrence

Hybrid Machine Translation (HMT)

- Integration of the statistical model with linguistic knowledge bases (Chiang, 2005; Marcu et al., 2006; Zollmann and Venugopal, 2006, …),

An RBMT approach to MWE processing: OpenLogos

OpenLogos system

(Scott, 2003; Scott and Barreiro, 2009; and Barreiro et al., forthcoming)

Linguistic Knowledge Base

Dictionaries (source, target and transfer)	Semantico-syntactic rules - analysis, transfer and generation	Semantic Tables **SEMTAB** - language-pair specific rules

An RBMT approach to MWE processing: OpenLogos

SAL - **S**emantico-syntactic **A**bstraction **L**anguage

- Taxonomy: 3 levels organized hierarchically: **Supersets** / **Sets** / **Subsets**

Semantico-Syntactic continuum from NL word to Word Class

• Literal word:	*airport*	
• Head morph:	*port*	
• **SAL Subset:**	**Agfunc**	(agentive functional location)
• **SAL Set:**	**func**	(functional location)
• **SAL Superset:**	**PL**	(place)
• Word Class:	N	

SAL combines both the lexical and the compositional approaches in order to process different types of MWE

An RBMT approach to MWE processing: OpenLogos

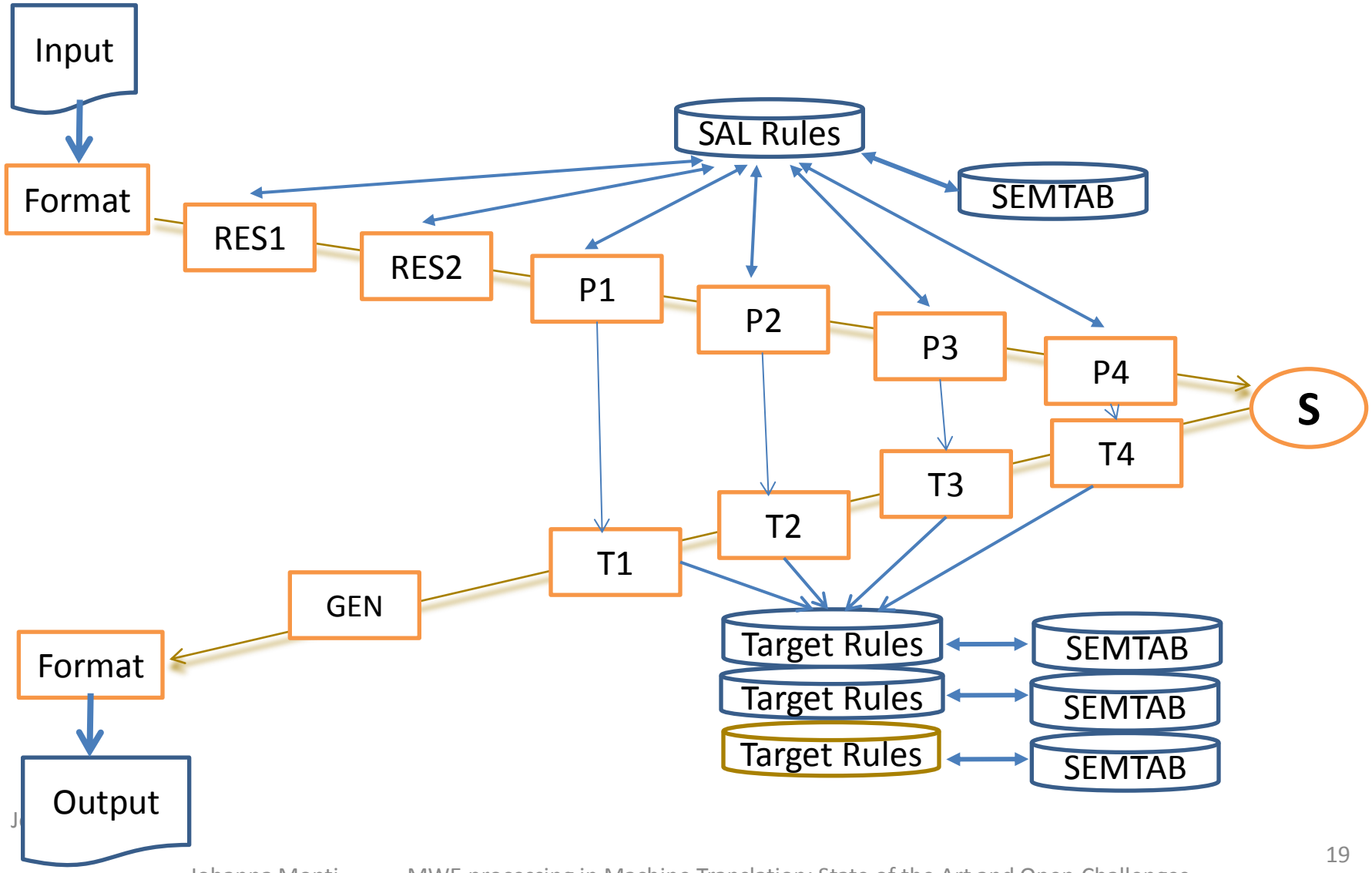
SEMTAB rules

- key component in MWE processing in OPENLOGOS
- able to handle different types of MWE
- invoked after dictionary look-up and during the execution of target transfer rules to solve analysis and lexical ambiguity problems
- identification, disambiguation and translation of MWE in context
- single deep-structure rules match multiple surface-structures and produce correct target transfers

Some examples:

- En. *mix up (VT) N (human) in* → It. *confondere N in*
- En. *mix up (VT) N (ingredient)* → It. *mescolare N*
- En. *mix up (VT) N (medicine)* → It. *preparare N*
- En. *mix up (VT) with* → It. *confondere con*
- En. *mix up (VT) N (human,info) with* → It. *confondere N con*

OpenLogos Architecture



MWE in SMT: state of the art

Word-for-word approach and the *noisy channel approach* (Brown et al., 1993)

- flat-structured models
- elements of syntax are captured implicitly
- Inability to handle many-to-many correspondences

Phrase-based translation model (Zens et al., 2002; Koehn et al., 2003) and the *maximum entropy approach* (Och and Ney, 2002).

- flat-structured model
- MWE marked and aligned as parts of consecutive phrases /not treated as special cases
- Able to handle many-to-many mappings
- Problems with discontinous MWE

Integration of PB-MT models with linguistic knowledge

- MRD and glossaries as phrases in the phrase-based table or to substitute unknown words
- Integration of syntactic and semantic structures

Phrase tables in SMT

" , per la gestione del presente ||| " for the management of this ||| 0.245841 0.000386953 0.245841 0.0788203 2.718 ||| ||| 1 1

" , per la gestione del ||| " for the management of ||| 0.245841 0.000632227 0.245841 0.0841843 2.718 ||| ||| 1 1

" , per la gestione ||| " for the management ||| 0.245841 0.00310736 0.245841 0.143357 2.718 ||| ||| 1 1

" , per la quale sono richiesti ||| " , requiring ||| 0.718868 3.33037e-08 0.718868 0.00289219 2.718 ||| ||| 4 4

" , per la ||| " for the ||| 0.0491683 0.00479878 0.245841 0.210926 2.718 ||| ||| 5 1

" , per ||| " for ||| 0.0491683 0.039521 0.245841 0.339868 2.718 ||| ||| 5 1

MT processing some recent approaches

Lambert & Banchs [2006]	• MWE should be identified and grouped with the corresponding translations prior to the alignment process
Wu et al. [2008]	• Construction of the *phrase tables* by means of bilingual dictionaries to improve SMT.
Zhixiang Ren et al. [2009]	• The integration of domain-specific bilingual MWE significantly improves translation.
Korkontzelos & Manandhar [2010]	• The knowledge about MWUs produces significant improvements
Bouamor et al. [2011]	• The integration of contiguous MWUs and their translation improves translation quality. .

Some good news from the MT community

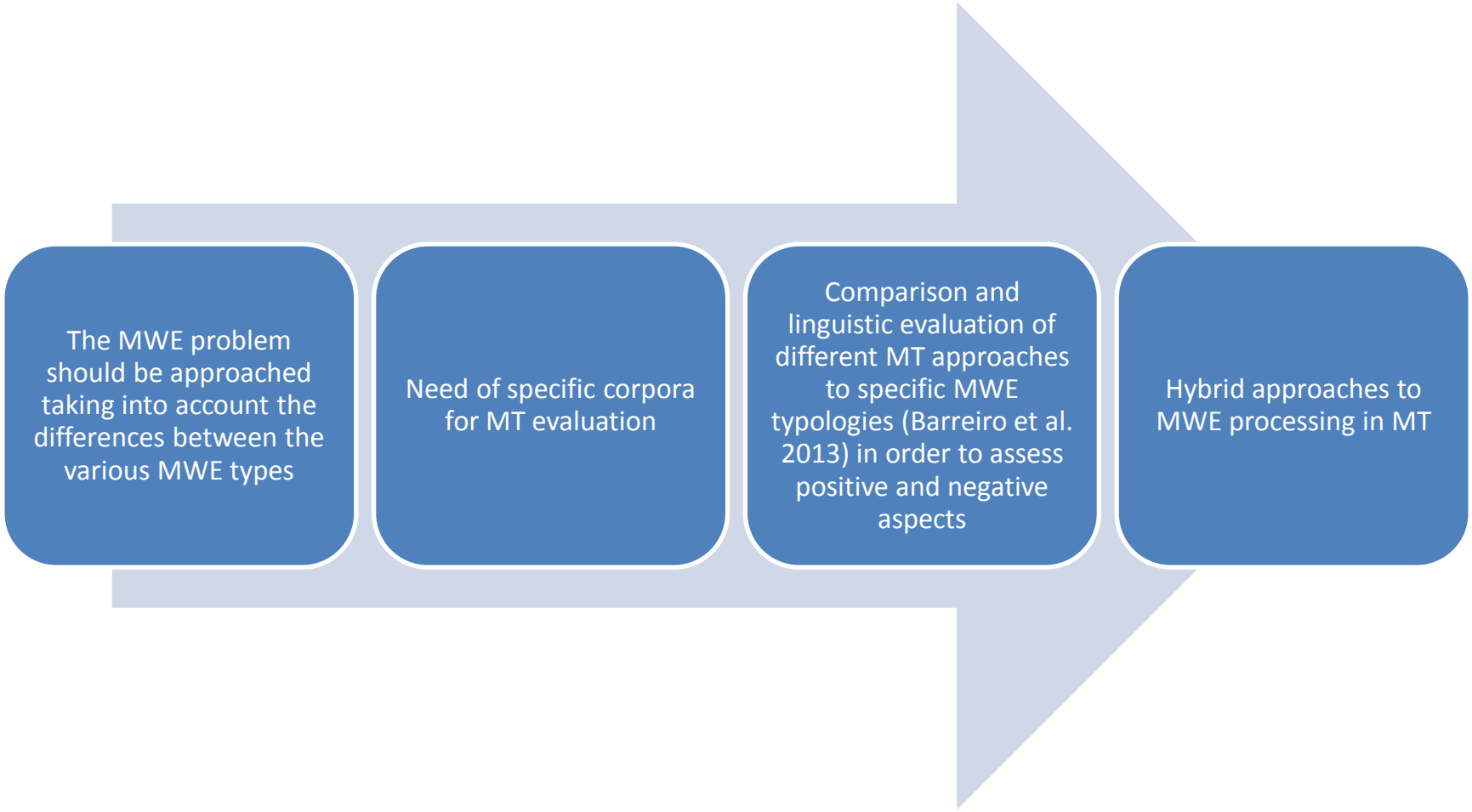
Growing attention to MWE processing in MT and Translation Technologies

Acknowledgement that it is not possible to create large-scale applications without properly handling MWEs of all kinds.

MTSummit 2013 workshop on ***Multi-word Units in Machine Translation and Translation Technology***

Closer interaction between NLP researchers, experts in phraseology (including computational phraseology) , terminologists and translation practitioners

Future challenges



Thank you for your kind attention!


Questions?