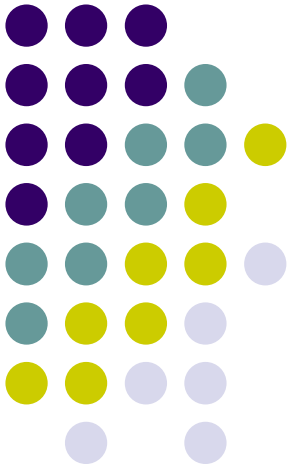
Consistency and completeness of multiword expressions during translation

Katerina Zdravkova
University Sts Cyril and Methodius, Skopje, Macedonia

Aleksandar Petrovski
International Slavic University, Sveti Nikole, Macedonia

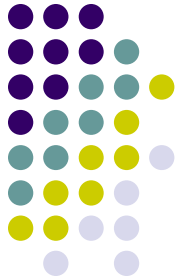
Tomaž Erjavec
Department of Intelligent Systems, Jožef Stefan Institute, Slovenia



Overview

- System for extraction of MWEs and their translations
- Evaluation of the system
- Main drawbacks
- Consistency of the system
- Completeness of the system
- Conclusion and further work

System for extraction of MWEs and their translations



- Extraction
- Syntactical filtering
- Translation and cross evaluation
- Evaluation of the results

MWEs extraction

- 3500 candidate MWEs, including some useless:
  - тоа би ја / *toa bi ja*, instead of тоа би ја усреќило / *toa bi ja usrekjilo* = that will make her happy
  - рече тој со / *reche toj so*, instead of рече тој со недоверба / *reche toj so nedoverba* = he said with mistrust;

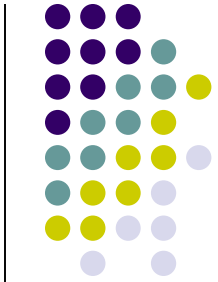
Syntactical filtering of MWEs

- Less than 500 phrases, sometimes inflections of the same phrase:
  - атомската бомба */ atomskata bomb* = the atomic bomb*,* атомски бомби */ atomski bombi* = atomic bombs
  - обичен човек */ obichen chovek* = an ordinary man, обичните луѓе */ obichnite lugje* = the ordinary men or the ordinary people,
  - шаховска табла */ shahovska tabla* = a chess board, шаховската табла */ shahovskata tabla* = the chess board

Translation and cross evaluation of MWEs

- Translation and cross evaluation
- This process extracted 968 English candidate MWEs
- They were translated using the system to:
  - Macedonian
  - Slovene
- and evaluated together

The learning corpus



- Multilingual corpus of Orwell's 1984 (created within Multext-East)
  - English
    - 6701 sentences
    - 104302 words
  - Macedonian
    - 6712 sentences
    - 98846 words
  - Slovene
    - 6684 Sentences
    - 90760 words

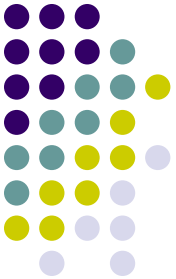
The crucial problems

- Manual translator of Orwell's 1984 was either inconsistent or had "an artistic freedom"
- Inflectional paradigms, which are richer in the Slavic languages can influence the translation
- The context in which the same target MWE appeared can also influence its translation

Incompleteness due to lexical inconsistency

Language	English	Macedonian	Slovene
Multiword expression	**the seconds were ticking by**	секундите минуваа (отчукувајќи)	sekunde so tiktakale mimo
Mac 1: секундите минуваа отчукувајќи ...			
Mac 2: секундите минуваа бескрајно долги ...			
Multiword expression	**almost on a level with**	речиси на исто (со)	no translation
Mac 1: ... речиси на исто ниво со ...			
Mac 2: ... речиси на исто рамниште со ...			
Slov 1: ... skoraj na ravni z ...			
Slov 2: ... skoraj v isti višini z ...			
Multiword expression	the first thing	(прва работа што мора) да ја сфатиш е	**prva stvar ki jo moraš ...**
Eng 1: the first thing for you to understand ...			
Eng 2: the first thing you must realize ...			

Incompleteness due to inflections



Language	English	Macedonian	Slovene
Multiword expression	**smell of her hair**	(мирисот) на нејзината коса	vonj njenih las
Mac 1: ... (пријатниот) мирис на нејзината коса			
Mac 2: мирисот на нејзината коса			
Multiword expression	**ideologically neutral**	идеолошки неутрален	ideološko nevtralen/na
Slov 1: ... (nobena beseda ... ni bila) ideološko nevtralna			
Slov 2: ... (predmet govora ni bil) ideološko nevtralen			
Multiword expression	**against us**	против нас	po robu (proti nam)
Slov 1: ... (nikdar ne) postavi po robu			
Slov 2: ... (in se nam) postavila po robu			

Incompleteness due to the context

Language	English	Macedonian	Slovene
Multiword expression	**for more than half an hour**	(за) повеќе од половина час	za več kot pol ure
Eng 1: ... and never for more than half an hour at a time Mac 1: ... и никогаш повеќе од половина час			
Eng 2: ... to turn off the telescreen for more than half an hour Mac 2: ... да го држат исклучен телекранот повеќе од половина час			
Multiword expression	**definitive edition**	дефинитивното издание	no translation (dokončna izdaja)
Eng 1: ... (the eleventh edition is the) definitive edition ... Slov 1: ... (enajsta izdaja je) dokončna			
Eng 2: ... (we were producing a) definitive edition ... Slov 2: ... (pripravljali smo) končno izdajo			
Mac 2: ... никогаш не работев во одделот за препишување Slov 2: ... (nikdar nisem bila v) prepisovalni ekipi Eng 2: (i was never in) the rewrite squad			

Consistency of the system

- Consistency measured with Herfindahl-Hirschman Index (*HHI*) measure (Itagaki, 2007)

$$HHI = \sum_{i=1}^{n} S_i^2$$

- *HHI* is applicable to multiword expressions, replacing the single words to lexical units
- Relative consistency:

$$RC = \frac{HHI}{|MWE|}$$

Completeness of the system

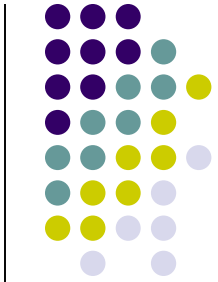
- We propose an index of completeness as its upgrading, to express the degree of correct translation of a complete MWE
- *DG* -> degree of generated MWE

$$DG = \frac{length(generated\ MWE)}{length(complete\ MWE)}$$

- *CC* -> combined completeness of the MWE

$$CC = \frac{1}{m}\sum_{j=1}^{m} S_j^2 DG_j^2$$

Consistency and completeness of our system

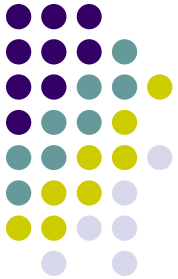


	Macedonian	Slovene
No translation	48	162
Partial inconsistency	127	91
HHI	836.75	778.25
RC	86.44%	80.40%
CC	83.42%	74.97%

Overview of our research

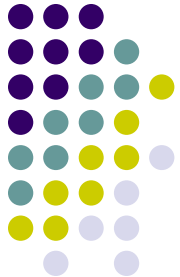
- We tried to define a framework for effective treatment of lexical units across languages
- We measured the consistency and the completeness of generated translations of MWEs existing in the small parallel Multext-East corpus

Extensions (1)



- Implement the same approach to bigger corpora
  - Example: the raw material obtained when Moses SMT toolkit, which was implemented over SETimes corpus (Slovene is not a part of this corpus)
- Incorporation of lexical entries +
- Extension with semi-fixed and flexible MWEs.
  - Decrease of inconsistency and incompleteness due to inflections

Extensions (2)



- Lexical cohesion
  - Extend the document-level translation to a larger collection
- Integration of the model into a hierarchical phrase-based SMT system
- Extraction of the common knowledge about multiword expressions out of a continuous context
  - Incorporation of acquired knowledge into a translation system capable to competently deal with MWEs

References

1. P. F. Brown et al. A statistical approach to machine translation, *Computational linguistics 16.2*, pp. 79-85, 1990.
2. M. Galley, C. D. Manning. Accurate non-hierarchical phrase-based translation, *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 966-974, 2010.
3. M. Itagaki, T. Aikawa, X. He. Automatic Validation of Terminology Translation Consistency with Statistical Method, *Proceedings of MT summit XI, 269-274*, pp. 269-274, 2007.
4. H. Caseli, C. Ramish, M. Nunes, A. Villavicencio. Alignment based extraction of multiword expressions, *Language Resources & Evaluation 44*, pp. 59-77, 2010.
5. R. Jackendoff. 'Twistin' the night away, *Language 73*, pp. 534-559, 1997.
6. J. Tiedemann. To cache or not to cache, Experiments with Adaptive Models in Statistical Machine Translation, *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp. 189-194, 2010.
7. Y. Zhang, V. Kordoni, A. Villavicencio, M. Idiart. Automated Multiword Expression Prediction for Grammar Engineering, *Proceedings of the 5th Workshop on Important Unresolved Matters*, pp. 44-52, 2006.
8. P. Koehn, F. J. Och, D. Marcu. Statistical phrase-based models, *Proceedings of NAACL 2003*, pp. 48-54, 2003

9. M. Stolikj, K. Zdravkova. Resources for Machine Translation of the Macedonian Language, *online Proceedings of ICT Innovations 2009*.
10. K. Zdravkova, A. Petrovski. System for extraction of potential multi-word expressions and prediction of their translations from a multilingual corpus, *PARSEME 2nd general meeting*, poster 43, 2014.
11. T. Erjavec. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages, *Language Resources and Evaluation*, Vol. 46 / 1, pp. 131-142, 2012.
12. V. Vojnovski, S. Džeroski, T. Erjavec. Learning PoS tagging from a tagged Macedonian text corpus, Proceedings of SiKDD 2005, Ljubljana, Slovenia, pp. 199-202, 2005.
13. L. Guillou. Analysing Lexical Consistency in Translation, *Proceedings of DiscoMT*, Sofia, Bulgaria, pp. 10-18, 2013.
14. A. Petrovski, K. Zdravkova. How to create a MWE lexical entry?. *PARSEME 3rd general meeting*, poster 8, group B, 2014.
15. G. Ben, D. Xiong, Z. Teng, Y. Lu, Q. Liu. Bilingual Lexical Cohesion Trigger Model for Document-Level Machine Translation, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 382-386, 2013.
16. S. Stymne, J. Tiedemann, C. Hardmeier, J. Nivre. Statistical Machine Translation with Readability Constraints, *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pp. 375-474, 2013.

Thank you for your attention

☺ ☺ ☺