TermWise

A Computer Assisted Translation Tool with Context-Sensitive Terminological Support

Kris Heylen*, Stephen Bond*, Dirk De Hertog*, Ivan Vulić†, Hendrik Kockaert* ‡

*QLVL - Linguistics Department (KU Leuven), †LIIR – Department of Computer Science (KU Leuven), ‡University of the Free State, Bloemfontein

Federale Overheidsdienst Justitie

PROJECT AIM: Corpus-based terminological support for specialised translators

ANALYSIS OF USER NEEDS :Lack of functionality in current CAT-tools

User Partner: Translation Service of the **Belgian Federal Justice Department**

- translates texts heavy with terminology and legal phraseology. (laws, reports,..)
- Both language versions of laws are authorative -> incorrect or inconsistent translation of terminology leads to legal uncertainty.
- Original documents are often mixtures of Dutch and French

Current commercial CAT-tool provides insufficient terminological support:

- Term Memory only finds previous translations on the segment level, not on the level of terms and expressions.
- Term Base is initially empty and needs time-consuming manual updates. Only a limited outdated version is currently available for Belgian legal terminology
- Concordancer only finds manually defined expressions in the Term Memory and does no sort previous occurrences for relevance to current assignment
- Only previous in-house translation are available in the Term Memory, not the large amount of Belgian legal documents online.

CASE STUDY: Belgian Legal Domain, Dutch-French Translation

SOLUTION OF USER NEEDS: **Term & Phrase Memory**

A separate module integrated in a CAT tool, with following functionality:

- Access to previous translations of subsegmental domain-specific expressions, single and multi-word
- Examples of usage in context to infer correct phraseology
- Information about the source documents of the translation example
- Examples from all relevant documents that are available online
- Sorting the examples by relevance to the current assignment
- Easy access to the examples from within the CAT-tool

The TermWise .project delivers **proof-of-concept** for:

- Language-independent automatic knowledge acquisition of bilingual terms and phraseology from large online corpora
- Server-Client archicture for a cloud-based Term & Phrase Memory
- Term & Phrase Memory functionality in case study of Belgian legal domain

RESEARCH: Automatic Knowledge Acquisition from Large Online Legal Corpora

INPUT: Parallel Dutch-French Legal Corpus, sentence aligned

- Belgisch Staatsblad/Moniteur Belgie 1997-2006, 100M (Van Allemeesch 2010)
- Enriched with meta-information (date, federal entity, ministry, department)

AUTOMATIC TERM EXTRACTION: separately for Dutch and French:

Terminological expressions in legal domain or of variable length (Kjær (2007) and includes noun phrases, verb phrases, formulaic sequences. Extraction aims at:

- n-grams of variable length (up to 8 words)
- no predefined language-specific POS patterns

Algorithm: 2 properties to identify relevant expressions among n-grams::

- External independence: Can n-gram occur in different contexts? -> Maximazation of frequency differences relative to the n-1 and n+1 grams in an n-gram expansion progression (Silva et al. 1999)
- Internal coherence: Do words co-occur in an informational unit? -> Mutual Information of the n-gram's POS-sequence

(Details: De Hertog 2014)

RESULT 649,602 n-grams for French and 639,865 n-grams for Dutch

BILINGUAL TERM ALIGNMENT: linking Dutch to French n-grams

Task Provide for each Dutch n-gram a ranked subset of likely translations from the French n-grams list and vv. High performance for low-frequency n-grams

Algorithm: SampLEX, adapted it to handle n-grams of variable length

- aligned sentences are represented as a bag-of-terms from NL&F n-gram lists
- Strategy of data reduction to for better performance on low-frequency terms
- Iterative sampling of sub-corpora and accumulative evidence for alignment
- benchmarked against other Bilingual Lexicon Extractions models

(Details: Vulić and Moens 2012).

RESULT: Dutch n-gram list with translation probabilities of French n-grams above (cut-off), and vice versa

sur la proposition du conseil d' administration	
op voorstel van de raad van bestuur	Prob: 0.621
op voordracht van de raad van bestuur	Prob: 0.379
16 mai 1989 et 11 juillet 1991	
16 mei 1989 en 11 juli 1991	Prob: 1.0
sur la proposition du ministre	
de voordracht van de minister	Prob: 0.481
op voorstel van de minister	Prob: 0.111
op voordracht van de minister	Prob: 0.074
...	...

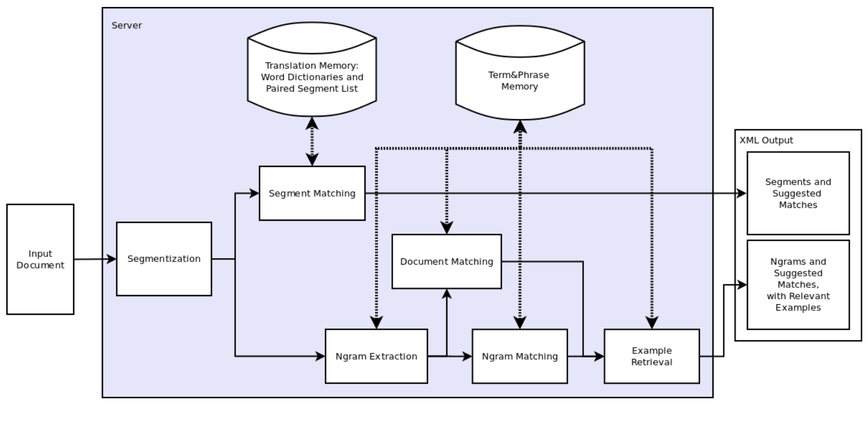
OUTPUT: Bilngual, aligned N-gram database with meta-data

Aligned Dutch-French n-gram lists+ document and sentence ID of each occurrence of a candidate translation-pair in the corpus together with meta-data of document

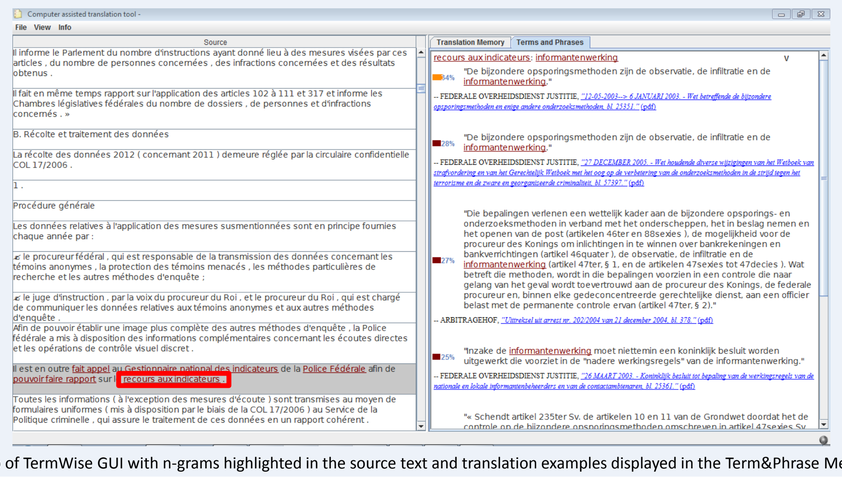
IMPLEMENTATION: User Interface with Context-sensitive Look-up of Terminology Translations

SERVER-CLIENT ARCHITECTURE

- User uploads a new translation assignment in client (CAT-tool)
- Document is uploaded to server and segmentized into sentences
- N-grams from database are detected in each segment
- 2 types of similarity calculations (bag-of-n-grams, cosine) on server:
  1. Segments matched with all sentences in Staatblad (~TM fuzzy match)
  2. Assignment's similarity with all documents in Staatsblad
- For identified n-grams: concordances retrieved from Staatsblad + meta-data of the document they occurred in.
- Concordances are sorted based on the document similarity to current assignment (relevancy) and categorized by translation (disambiguation)
- Output from server sent as XML-file back to client



USER INTERFACE: Demo of functionality in Java

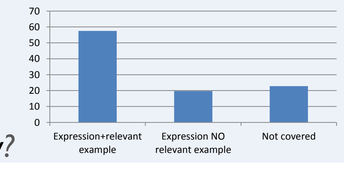


Screen cap of TermWise GUI with n-grams highlighted in the source text and translation examples displayed in the Term&Phrase Memory pane

USER EVALUATION: Coverage of terminology look-up needs

- 19 Students of Legal Translation, KULeuven@TMA
- Translators at Federal Justice Department (Feb. 2014)

For expressions whose previous translations you would like to look up, how many were covered by the Term&Phrase Memory?



REFERENCES

Dirk De Hertog. 2014. TermWise Xtract: Automatic Term Extraction applied to the legal domain. Phd, KU Leuven.

Anne Lise Kjær. 2007. Phrasemes in legal texts. In Harald Burger, Dmitrij Dobrovol'skij, Peter K¨uhn, and Neal Norrick, editors, *Phraseology An International Handbook of Contemporary Research*, pages 506–516. WdG.

Joaquim Ferreira da Silva, Gael Dias, Sylvie Guillore, and Jose Gabriel Pereira Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and noncontiguous multiword lexical units. In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence* (EPIA), pages 113–132..

Tom Vanallemeersch. 2010. Belgisch staatsblad corpus: Retrieving French-Dutch sentences from official documents. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)

Ivan Vulić and Marie-Francine Moens. 2012. Sub-corpora sampling with an application to bilingual lexicon extraction. In Proceedings of the 24th International Conference on Computational Linguistics (COLING), pages 2721–2738..

Contact:: {kris.heylen, stephen.bond, dirk.dehertog, hendrik.kockaert}@arts.kuleuven.be, ivan.vulic@cs.kuleuven.be

Research funded by KU Leuven IOF grant KP/09/001. Special thanks to Martine Perpet of the FOD Justitie