SAF€

29.09.2015.

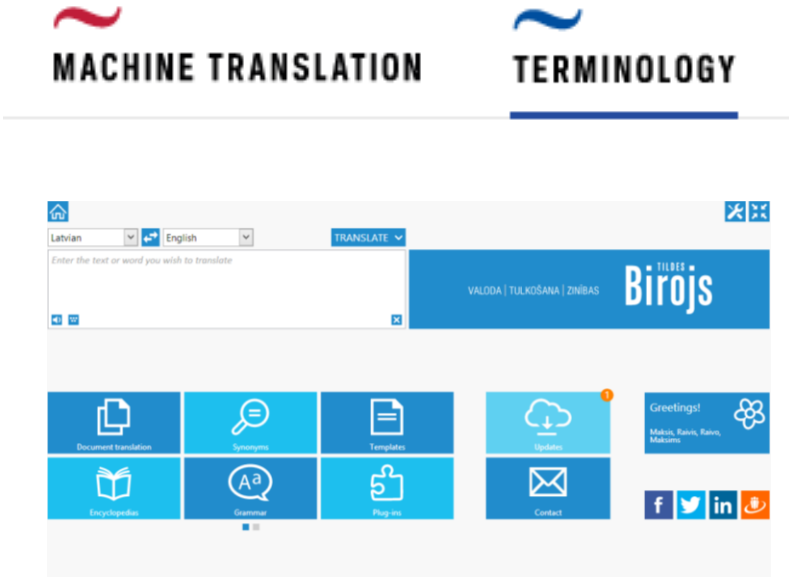
SIA Tilde

Project manager: Artūrs Vasiļevskis
E-mail: arturs.vasilevskis@tilde.lv

About Tilde

- Leading European language technology company
- Founded in 1991 in Latvia
- 135 employees in Tallinn, Riga, and Vilnius
- 5 staff PhDs
- 80+ scientific publications

Language Technology Products



- Machine translation
- Terminology solutions
- Proofing tools
- Mobile apps
- Online encyclopedia
- Digital information collection

Our Expertise

- EU leader in developing technology for smaller languages
- European Commission technology partner
- Collaboration with the world's leading universities

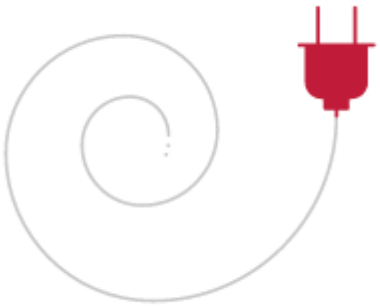
Machine Translation (MT)

Instant translation of texts, documents, and websites.

Tilde MT

Tailored to the specific translation needs of each organization, using their own style, terminology, and language.

MT integration



- Customer service forums
- E-commerce platforms
- Data analysis software
- Corporate intranet
- Mobile apps
- Live chats
- Websites

About SAF€ project

This project aims to demonstrate combination of **machine translation** and **semantic sentiment analysis** to capture the "wisdom of the crowd" for financial decision making.

Project duration:

01/09/2013

to

03/09/2015

(25M)

eurostars™ SAF€

Consortium

- **Semlab (Netherlands):**
  - Ontology development and sentiment analyses
  - processing and storage of translated social media sentiment

- **HWC Communications Limited (United Kingdom):**
  - social media data firehose provider
  - processing and storage social media data

- **TILDE (Latvia):**
  - Statistical Machine Translation
  - processing and storage of translated social media

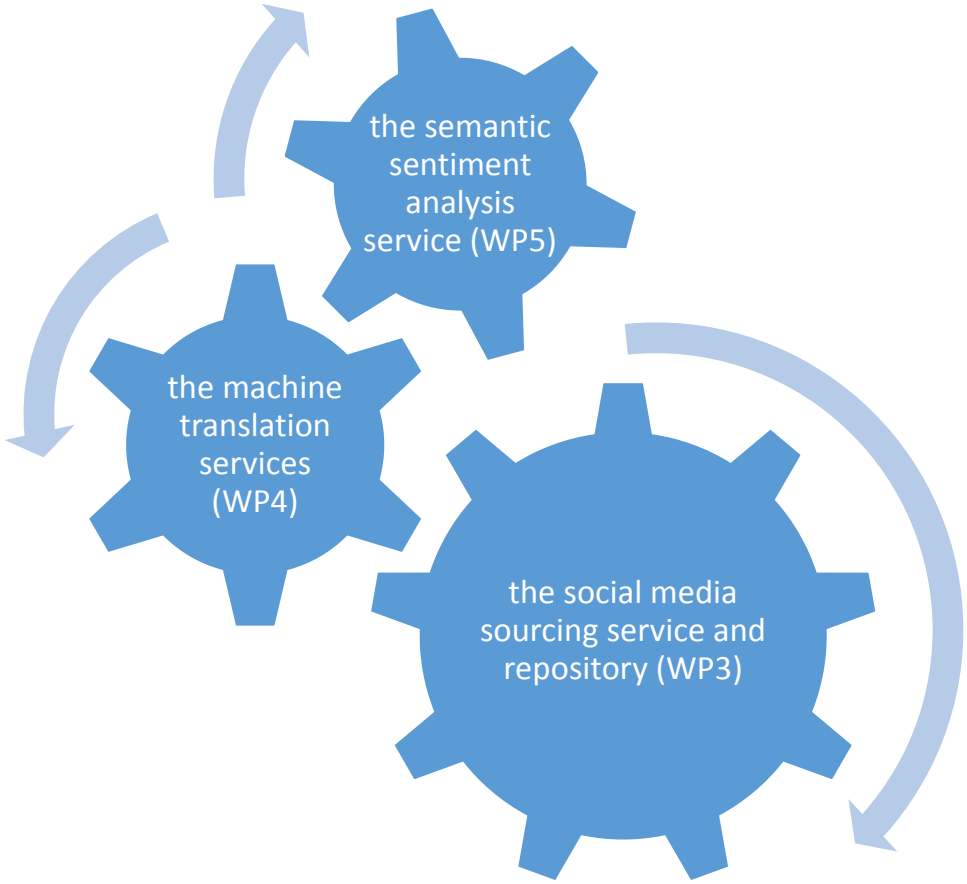
- **JRC Capital Management Consultancy & Research GmbH (Germany):**
  - functional requirements
  - exploitation and dissemination

eurostars™ SAF€

WP of the project

- WP1 - Prototype development : SemLab
- WP2 – Performance assessment : JRC
- WP3 – Web data sourcing and processing : HW
- WP4 - Machine Translation: **TILDE**
- WP5 – Sentiment Analysis : Semlab
- WP6 - Dissemination & Exploitation : JRC
- WP7 - Project Management - SemLab

Main building blocks



Tilde main contribution

- Tilde as Machine translation (MT) technology expert in the project leads development of MT systems, provides know-how and management of MT related tasks.

Tilde Tasks in SAFE

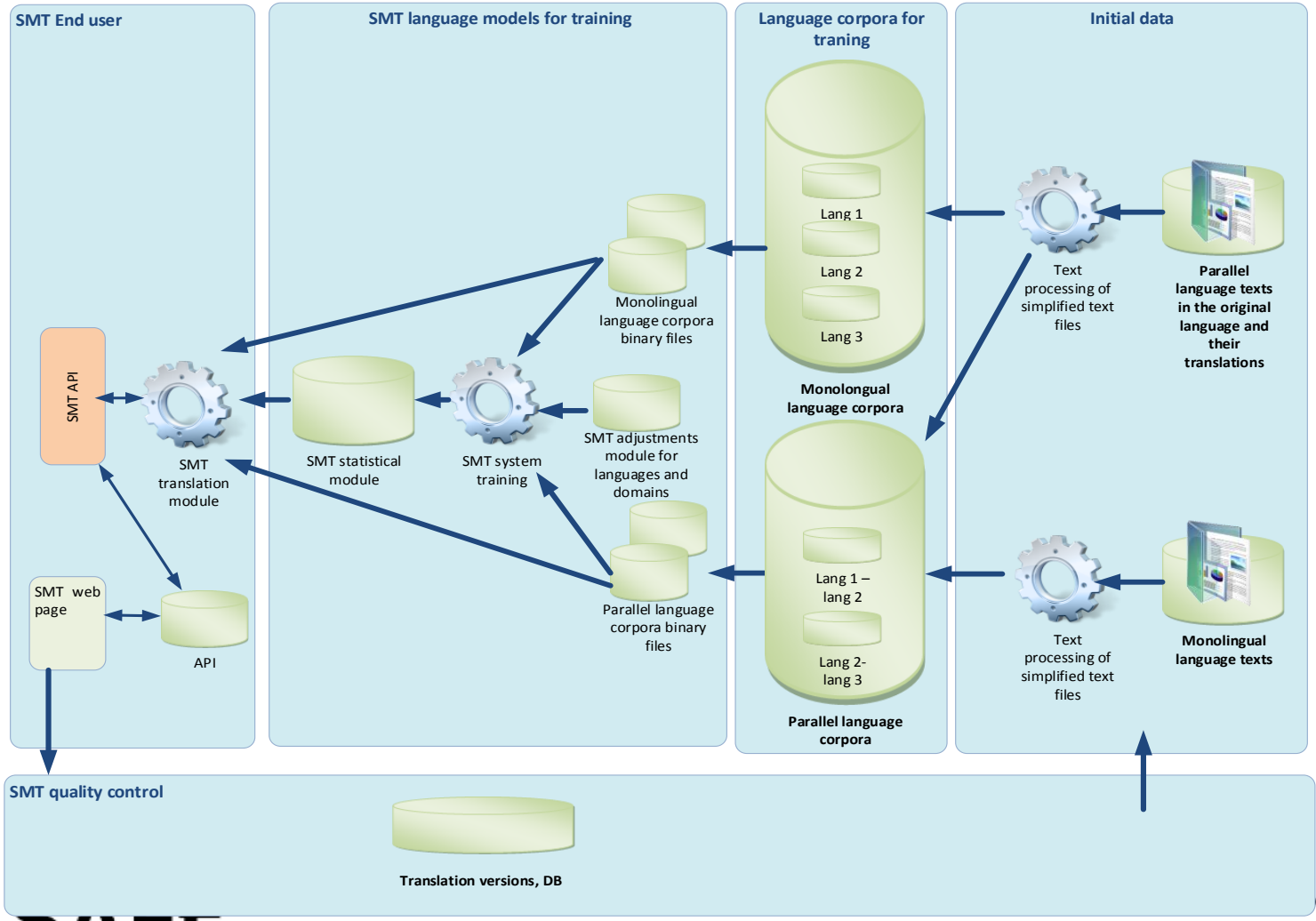
WP1	Tilde is involved in tasks to determine requirements and solution concepts for the prototype development and deployment.
WP2	Tilde involvement is in tasks related to assessment of the prototype from perspective of usage of machine translation subsystems, the quality will be determined using non-adapted SMT baselines and measuring improvement in BLEU scores.
WP3	Tilde is involved in preparation of the requirements for the data that will be collected, processed and used for SMT training.
WP4	To provide multilingual support the LetsMT! platform is used. The LetsMT! allows building (training) a statistical machine translation (SMT) system for any language pair if large enough data (parallel corpora and monolingual corpora) of good quality for particular language pair is provided. The machine translation system created with LetsMT! tools then can be easily accessed through API and integrated into the proposed platform. Although the LetsMT! platform could be used to build SMT system for any language pair, in this project as proof of concept we will focus on: : English, French, German, Polish, Dutch, Latvian, Swedish
WP5	-
WP6	Tilde will be involved in preparation of Dissemination strategy for the project dissemination activities and implementation of activities in accordance agreed strategy.
WP7	-

eurostars™ SAFE

MT Challenges

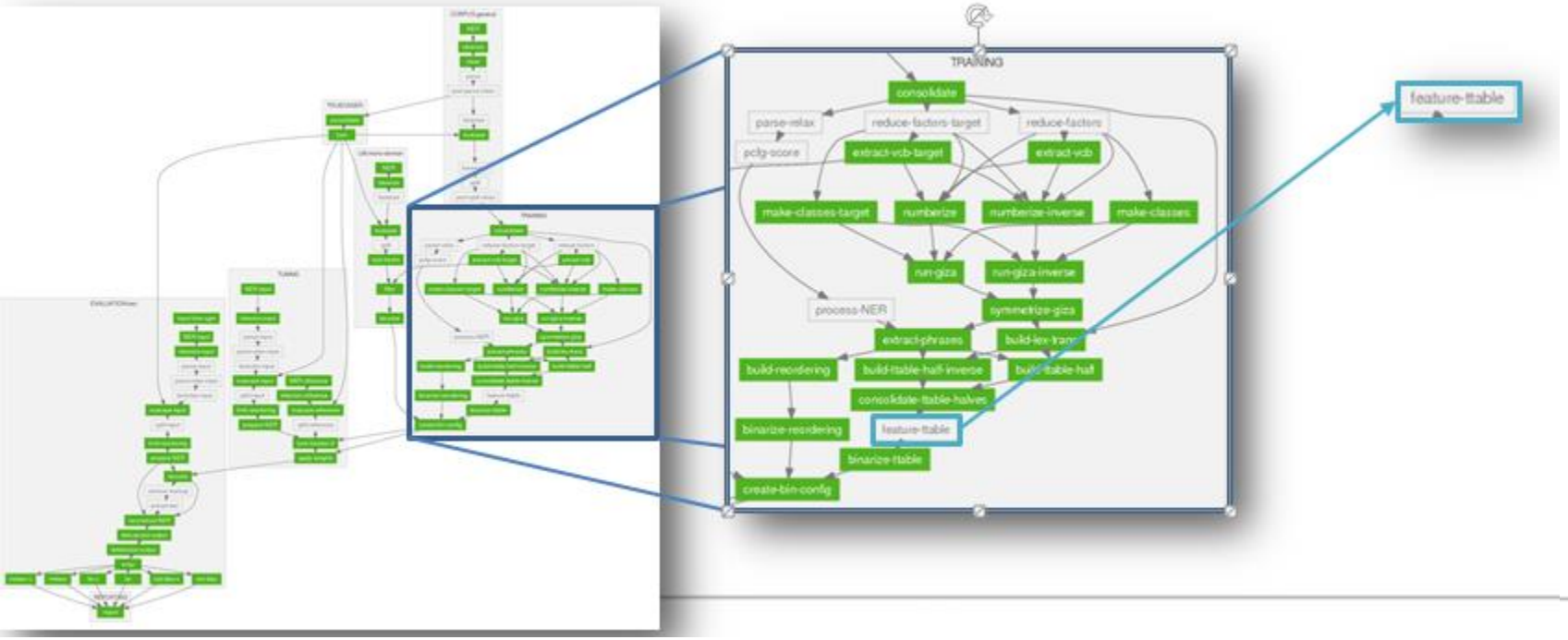
- Provide MT infrastructure for project specification
- Domain and channel specific (representative) parallel and monolingual data processing
- MT adaptation to specialized domains, domain specific key word and named entity recognition
- Integration of named entity recognition tools
- Specific key words and glossaries (terminology) integration in MT to create a new type of specialized MT engine for use in other WP
- Develop specialised MT systems

AWS infrastructure



Translation model adaptation

- A new module in the training workflow
- Experimental status - has to be enabled manually by LetsMT/Moses developers through a command line interface
- Example: the German - English system training chart:



Main tasks

- Evaluation of requirements for the LetsMT adaptation procedures
- Data processing and enrichment
- SMT systems development
- SMT system evaluation

Data requirements

- EN mono in-domain specific corpora (500M segments)
- Development-test set – 2000 sentences in all project languages (EN-xx)
- Some Twitter-like parallel corpora for project languages
- Lists of domain-specific abbreviations and their translations, both EN and project languages, mono & parallel, eg.
  - hrs — hours — en
  - val. — valūta — currency — lv-en
  - pēd.c. — pēdējā cena — last price — lv-en
- Company names, products, stock index lists
- Lists of non-translatable tokens (NTT)
- List of organizations and their hashtags

eurostars™ SAF€

Social Network specific in SMT

- Recognizes non-translatable tags and leave as is in the translation
  - Hashtag #UN, #CWC15, #Fact, #Newsmax
  - Cashtag $AAPL, $MSFT, $TSLA, $BLDP, $PVSP, $CGRA
  - Mention @tim_cook, @hk_riga, @Ankit_Bera
  - Plustag +LielaKeda
  - Smileys :) (: ;) :-) :D ;D :P :p :* ;* ^_^ ^^ =] =) (= [= :] [: ;] [; :O :o O: o: :S :s :( ): ;(:@ :| -_-)=]= =(=[:[ ]: ;[ ]; :| (;
  - Retweet construction RT @Ankit_Bera:

SMT systems list trained for SAFE project

- Latvian – English
- Swedish-English
- German-English
- Dutch – English
- Polish- English
- French - English

The ViewerPro web client offers an at-a-glance overview of the current news in connected news sources. The application displays headlines and full text of all messages in a single inbox and provides intuitive visualizations to allow in-depth analysis of recent news events.

The navigation menu provides acces to the various functionalities of the application. Also the number of unread news items from all connected sources (social media, blogs, newswire etc.) are listed in the red balloon.

This section shows intuitive positive/negative sentiment graphs. In addition to a general graph showing the sentiment distribution over all events, several sub graphs are availabel. The sub graphs are arranged according to the main event topics (M&A, Value, Finace and Operations) and provide an overveiw of the event calsses in these topics.

Clicking a graph or an item in the list filters the page to reload the page, showing only data about this specific item. This provides a simple and intuitive way to drill down to the level of information required for your analysis.

In this section several lists are provided in which the subjects and events types found in a selection of news items are ranked according to the sentiment values found. Each list shows the ten items associated with the most positive sentiment and the ten most negatively rated items.

The top-10 lists are provided for all news item attributes including subject, event type, sources, author etc.

SemLab online sentiment analysis solutions

- newstape.semlab.nl
- http://www.newssentiment.eu

Questions?