Ontology-based translation memory maintenance

Andraž Repar[1,2], Senja Pollak[3]

[1] International Postgraduate School, Ljubljana, Slovenia
[2] Iolar, Ljubljana, Slovenia
[3] Jozef Stefan Institute, Ljubljana, Slovenia

Translation industry

According to SDL, *'Inconsistencies in the use of terminology' is the top cause of translation rework*

Iolar started developing its own terminology management system in cooperation with the JSI Institute:

- Terminology management
- (Bilingual) terminology extraction
- Definition extraction
- **Domain extraction**
- Good example extraction (Repar, Pollak, 2017)

Outline of the talk

- Introduction - What is a translation memory
- Text mining approach
  - Dataset
  - Experiments
- Evaluation and results
- Conclusion and future work

What is a translation memory

- Database of translations
- Standard in the translation industry
- Workflow: segment-translate-save

Why?

- Leveraging translation memories
- Reuse for e.g. machine translation

Type	Segments	Words	Characters	Percent	Placeables	Tags
PerfectMatch	0	0	0	0.00%	0	0
Context Match	0	0	0	0.00%	0	0
Repetitions	1	6	19	33.33%	0	0
Cross-file Repetitions	0	0	0	0.00%	0	0
100%	0	0	0	0.00%	0	0
95% - 99%	0	0	0	0.00%	0	0
85% - 94%	0	0	0	0.00%	0	0
75% - 84%	0	0	0	0.00%	0	0
50% - 74%	0	0	0	0.00%	0	0
Internal:						
95% - 99%	0	0	0	0.00%	0	0
85% - 94%	1	6	22	33.33%	0	0
75% - 84%	0	0	0	0.00%	0	0
50% - 74%	0	0	0	0.00%	0	0
New	1	6	19	33.33%	0	0
Total	3	18	60	100%	0	0

Translation memory - an example

Translation Memory eXchange (TMX) file format

```
<tu creationdate="20140527T094016Z" creationid="AR_00" changedate="20140527T094016Z" changeid="AR_00" lastusagedate
    ="20140527T094016Z">
  <prop type="x-Origin">TM</prop>
  <prop type="x-OriginalFormat">TradosTranslatorsWorkbench</prop>
  <tuv xml:lang="en-GB">
    <seg>The Subfund remained invested almost completely in USD-denominated corporate bonds with a short to medium residual
        term.</seg>
  </tuv>
  <tuv xml:lang="sl-SI">
    <seg>Podsklad je skoraj v celoti ostal vložen v dolarske podjetniške obveznice s kratko do srednje dolgo zapadlostjo
        .</seg>
  </tuv>
</tu>
<tu creationdate="20140527T100115Z" creationid="AR_00" changedate="20140527T100115Z" changeid="AR_00" lastusagedate
    ="20140527T100115Z">
  <prop type="x-Origin">TM</prop>
  <prop type="x-OriginalFormat">TradosTranslatorsWorkbench</prop>
  <tuv xml:lang="en-GB">
    <seg>The focus was on issuers from the IT and consumer goods sectors.</seg>
  </tuv>
  <tuv xml:lang="sl-SI">
    <seg>Usmeril se je na izdajatelje iz sektorjev informacijskih tehnologij in potrošniških dobrin.</seg>
  </tuv>
</tu>
```

Why the need for maintenance

- Old technology - since the beginning of 1990s
- (Normally) one translation memory per project/domain
- Lots of users over a long period of time - resulting in:
    - Wrong languages in the translation memory
    - Out-of-domain content in the translation memory
- Initial domain too broad for modern applications (e.g. machine translation, terminology extraction)

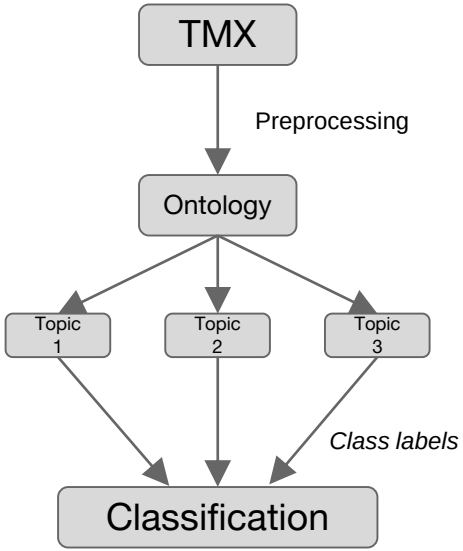
Dataset

MarLegFin translation memory

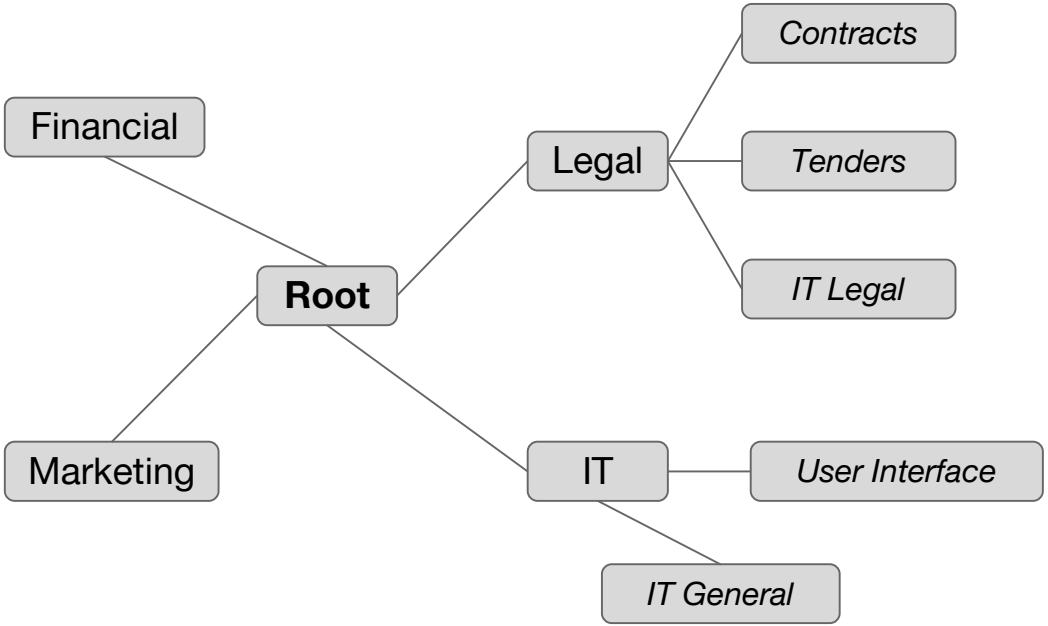
- Created sometime after 2000
- Almost 250,000 EN-SL segments
- Marketing, Legal, Finance
- However, other domains can also be found
- No way of manually extracting individual domains

Experiments

- Import data into OntoGen (Fortuna et. al., 2007)
  - One segment is one document ---- difficult to classify!
  - Removed all segments with less than 8 words
  - K-means clustering + manual grouping into topics
  - Manual evaluation of extracted topics
- Use the extracted topics as a shortcut for classification of new content
  - Topics as class labels, StringToWordVector
  - Test performance of Naive Bayes Multinomial, SVM and J48 in Weka (Hall et.



Results


Financial
Root
Marketing
Legal
Contracts
Tenders
IT Legal
IT
User Interface
IT General

Manual evaluation of OntoGen results

50 segments per topic

Topic	*Precision*
Financial	0.76
IT General	0.80
IT User Interface	0.86
Legal Contracts	0.80
Legal IT Legal	0.86
Legal Tenders	0.78
Marketing	0.80
Average	**0.81**

Classifier performance - Main topics

Majority class accuracy = 0.406

	J48	SMO	NB Multinomial
Accuracy	0.597	0.619	0.671
Precision	0.615	0.608	0.678
Recall	0.597	0.619	0.671
F-measure	0.576	0.610	0.673

Conclusion and future work

- OntoGen can be used for translation memory maintenance
- Classification performance was promising but not yet good enough

Future work

- Increase frequency of domain terminology
- Segment grouping based on translation memory metadata