Compound dictionary extraction and Wordnet A dangerous liaison?

Carla Parra Escartín, Héctor Martínez Alonso
University of Bergen, University of Copenhagen
PARSEME WG3

Object of study

- We focus on ways of automatically retrieving compound dictionaries from sentence-aligned corpora using WordNet for the pair of languages German→Spanish.
- German→Spanish compound correspondences are of the type 1:n:

(1)

Warm	*Wasser*	*Bereitung*
caliente	agua	preparación
warm	water	production

[ES]: ‘Preparación de agua caliente’

[EN]: ‘Warm water production’

(2)

Wärme	*Rückgewinnung*	*s*	*Systeme*
calor	recuperación	Ø	sistemas
heat	recovery	Ø	Systems

[ES]: ‘sistemas de recuperación de calor’

[EN]: ‘heat recovery systems’

- The ultimate aim is to integrate the extracted compound dictionaries in Statistical Machine Translation (SMT) tasks.

Gold Standard

Our Gold Standard consists of 168 compounds and their translations:

- They were extracted from the TRIS corpus [1], a specialised German→Spanish corpus.
- All compounds were split and tagged with their corresponding Part-of-Speech (PoS) tags [2].
- All translation correspondences were also PoS tagged [2].
- If a compound had several translation correspondences, each was stored as a different entry in the Gold Standard.

Compound-phrase matching

1. Given a split German compound C, there is a list of lemmas $C = [c_0, ..., c_n]$.
2. Given a Spanish sentence aligned to the German sentence that contains C, there is a list of lemmas $S = [s_0, ..., s_n]$.
3. Be $type(x)$ a function that retrieves the semantic type of a word, obtained from Wordnet.
4. For each German compound, Spanish sentence pair (C,S):
   (a) Locate the translated root of C in S by finding a lemma s_x in S with a semantic type that matches the root of the compound, i.e. $type(s_x) = type(c_n)$.
   (b) Locate the rightmost word in the Spanish phrase that translates C by finding a lemma s_y in S with a semantic type that matches the first lemma of the compound, i.e. $type(s_y) = type(c_0)$.
   (c) The candidate Spanish phrase that translates C is the span of words defined as $[s_x, ..., s_y]$.

Challenges faced

- PoS taggers: More damaging on the Spanish side when not locating phrase roots.
- WordNet coverage.
- Manual semantic matching:
  GermaNet has a potentially useful adjective classification that maps unevenly to the Spanish WordNet.

Acknowledgements

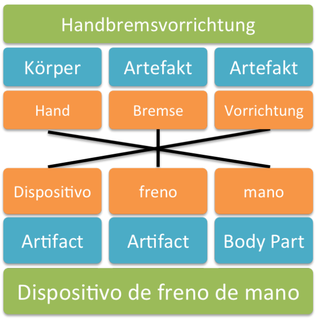
KØBENHAVNS UNIVERSITET

The current research was financed by the EU under the Marie Curie Actions, FP7 People programme (grant agreement 238405).

CLARA MARIE CURIE ACTIONS PEOPLE

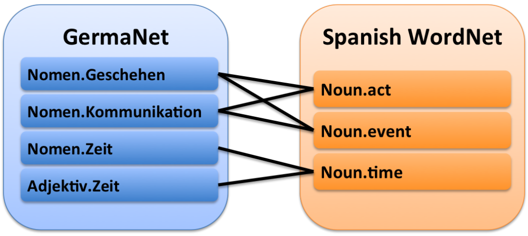
Working hypothesis: semantic types mapping

Our working hypothesis is that different formants of a compositional compound will share semantic features with their corresponding translational equivalents:



DE → ES semantic matching

- The semantic type matching had to be done manually.
- There are n:n and n:1 correspondences because GermaNet and the Spanish Wordnet do not share a common list of semantic types:



Matching metrics

We tested whether our hypothesis held for our Gold Standard:

	Number of items	Percentage
Total Pairs	168	100%
Perfect coverage pairs	93	55%
Perfect coverage German	46	27%
Perfect coverage Spanish	13	8%
WN coverage error on both	16	10%
Missing German roots	18	11%
Missing Spanish roots	19	11%

Conclusion and future work

- Expand the Gold Standard.
- Evaluate the PoS tagger and identify sources of error that might be avoided. Eventually test other PoS taggers.
- Redefine the $type(x)$ function to make it not only dependent on the first listed sense of each WordNet.
- Align semantic classes automatically using word-alignment techniques, or using the English WordNet as a pivot.
- Use supervised machine learning to predict Spanish phrase spans from the German compounds.

References

[1] Carla Parra Escartín, *Design and compilation of a specialized Spanish-German parallel corpus*, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (Istanbul, Turkey), European Language Resources Association (ELRA), May 2012, pp. 2199–2206 (English).

[2] Helmut Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees*, International Conference on New Methods in Language Processing (Manchester, UK), 1994, pp. 44–49.