DutchSemCor: Targeting the ideal sense-tagged corpus

Piek Vossen[1], Attila Görög[1], Rubén Izquierdo[2], Antal van den Bosch[2]

1 VU University Amsterdam. Amsterdam. The Netherlands
2 Tilburg University. Tilburg. The Netherlands

Build a sense-tagged corpus for Dutch

- **Represent all senses of words**
- **Represent the variety of contexts**
- **Provide information on the sense distribution**

- **3000 most frequent Dutch words (also most polysemous)**
- **100 examples per sense**
- **3-4 avg. senses per word --> 1 million tokens**

Sense tagged corpus

- **Annotation method**
  - **Sequential tagging**
  - **Targeted tagging**
- **Textual coverage**
  - **All - words**
  - **Lexical sample**
    - **Balanced-sense**
    - **Balanced-context**
    - **Sense-Probability**

Project Methodology. 3 Phases

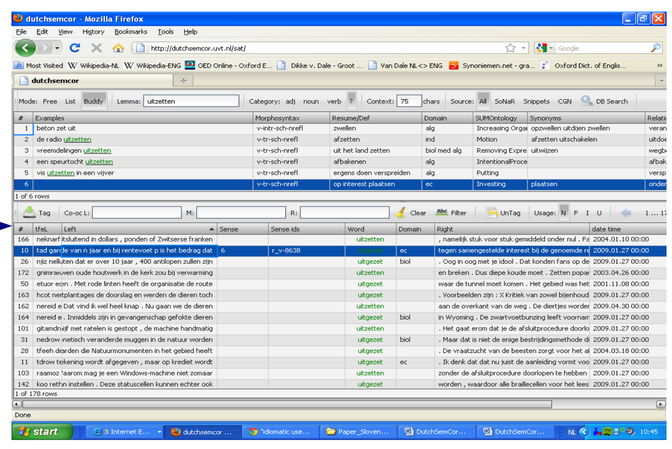
1)Human manual annotation

human + human

- Double annotation for each token
  - 282,503 tokens
  - Very clear and good examples selected

Goal: *Balanced-sense corpus*

SAT Tool - KWIC



67 %

5 %

28 %

A 500-million-token corpus is not big enough to create a balanced-sense corpus !!!

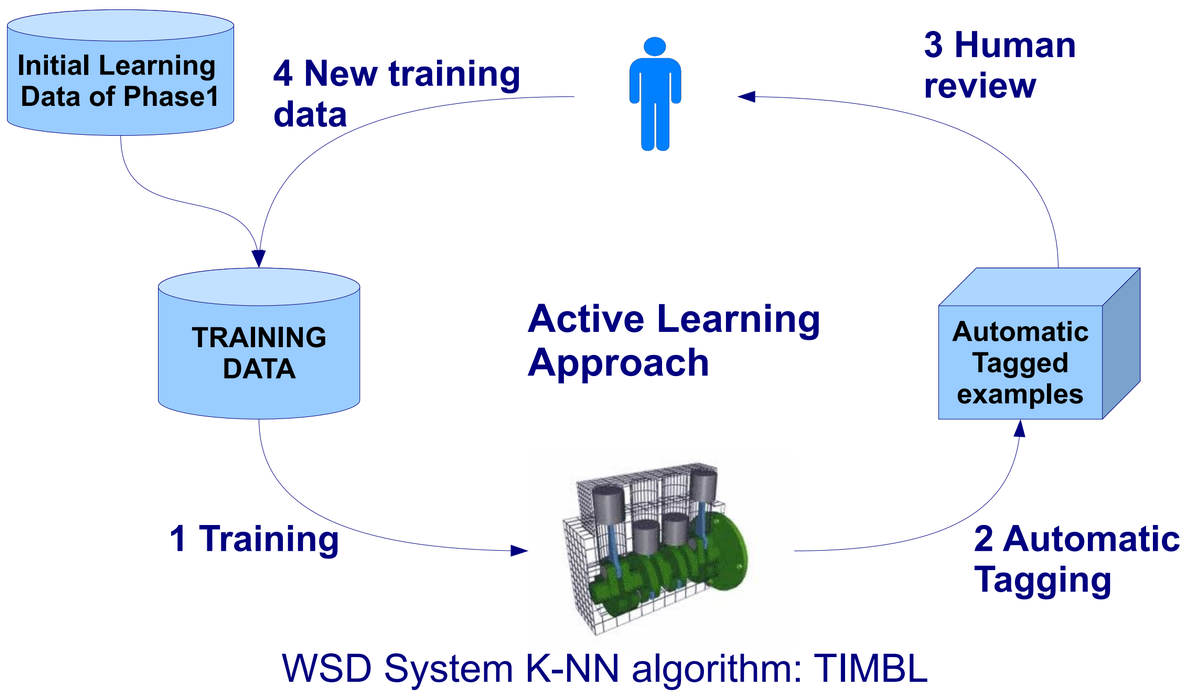
- Double annotation for each token
  - 282,503 tokens
  - Very clear and good examples selected
- At the end of the annotation
  - 80% senses with 25 of more examples
  - 90% of lemmas with 25 examples per sense
  - This set is called INITIAL LEARNING

2)Active Learning

human + computer

- Only lemmas performing < 80 in accuracy are processed
- 50 examples per sense according to:
  - TIMBL confidence
  - Distance to the nearest neighbor
    - Low Distance (LD): similar examples
    - High Distance (HD): different examples

Goal: *Balanced-context corpus*



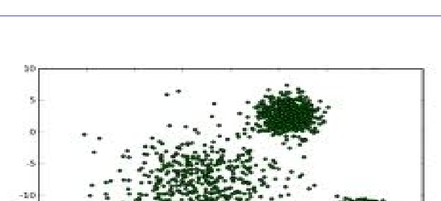
WSD System K-NN algorithm: TIMBL

Data	Token Accuracy	# Examples
Initial Learning	81.62	8641
IL + LowDist	78.87	13266
IL + LowDist_agree	85.02	11405
IL + HighDist	76.24	19055
IL + HighDist_agree	83,77	13359
IL + LowDist_agree + HighDist_agree	85.33	16123

3) Clustering

Goal: *sense-probability corpus*

- Similar to Word Sense Induction
- Clustering techniques different to WSD
- Cluster remain not tagged SONAR to:
  - Discover new senses
  - Use annotated instances to discover clusters with a predominant word sense and automatically tag the cluster



vrije Universiteit amsterdam

Tilburg University