

NCLT/CNGL  
Internal Workshop

24 July 2008

Ankit Kumar  
Srivastava

# LEARNING WORD TRANSLATIONS

NON PARALLEL  
CORPORA

Does syntactic context fare  
better than positional context?

# Learning a Translation Lexicon from non Parallel Corpora

- ❖ Motivation
- ❖ Methodology
- ❖ Implementation
- ❖ Experiments
- ❖ Conclusion

Master's Project

AT

University of  
Washington  
Seattle, USA

JUNE 2008

# { lexicon }

- Word – to – word mapping between 2 languages
- Invaluable resource in multilingual applications like CLIR, CL resource, CALL, etc.

Wahl	
election	0.85
ballot	0.10
option	0.02
selection	0.02
choice	0.01

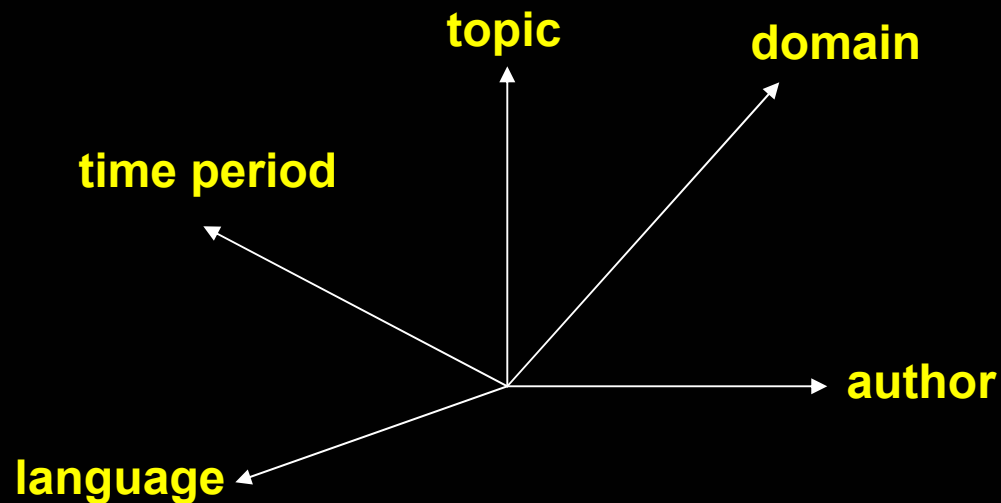
**Sheridan & Ballerini 1996**  
**McCarley 1999**

**Yarowsky & Ngai 2001**  
**Cucerzan & Yarowsky 2002**

**Nerbonne et al. 1997**

## { corpora }

- ◈ Parallel, comparable, non-comparable text
- ◈ More monolingual text than bitext
- ◈ 5 dimensions of nonparallelness
- ◈ Most statistical clues no longer applicable



# MOTIVATION

{ task }

Given any two pieces of text  
in any two languages...



...Can we extract word  
translations?

## { insight }

- ❖ If two words are mutual translations, then their more frequent collocates (context window) are likely to be mutual translations as well.
- ❖ Counting co occurrences within a window of size N is less precise than counting co occurrences within local syntactic contexts [**Harris 1985**].
- ❖ 2 types of context windows – Positional (window size 4) and Syntactic (head, dependent)

# { context }

Vinken will join the board as a nonexecutive director Nov 29 .

## POSITIONAL:

Vinken will join the board as a **nonexecutive** director Nov 29 .

## SYNTACTIC:

Vinken will join the board as a **nonexecutive** director Nov 29 .



# { algorithm }

- ◆ For each unknown word in the SL & TL, define the **context** in which that word occurs.
- ◆ Using an initial **seed** lexicon, translate as many source context words into the target language.
- ◆ Use a **similarity** metric to compute the translation of each unknown source word. It will be the target word with the most similar context.

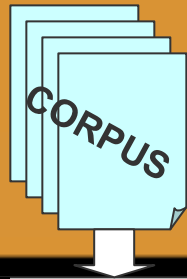
**Rapp 1995, 1999**

**Fung & Yee 1998**

**Koehn & Knight 2002**

**Otero & Campos 2005**

# IMPLEMENTATION



{ system }

**1** CORPUS  
CLEANING



**2** PCFG  
PARSING

# { pre-process }

## 1 Raw Text Corpora:

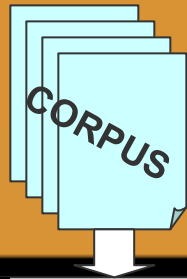
	ENGLISH	GERMAN
DATA	Wall Street Journal (WSJ)	Deutsche Presse Agentur (DPA)
YEARS	1990,1991 and 1992	1995 and 1996
COVERAGE	446 days of news text	530 days of news text

## 2 Phrase Structures:

Stanford Parser (Lexicalized PCFG) for English and German  
<http://nlp.stanford.edu/software/lex-parser.shtml>

[Klein & Manning 2003]

# IMPLEMENTATION



{ system }

**1** CORPUS  
CLEANING

**2** PCFG  
PARSING

**3** PS TO DS  
CONVERSION

**4** DATA  
SETS

July 24, 2008

Lexicon Extraction ~ Ankit

11

# { pre-process }

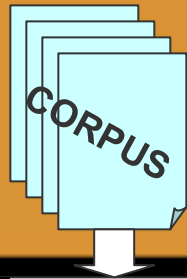
## 3 Dependency Structures:

Head Percolation Table [**Magerman 1995; Collins 1997**] was used to extract head-dependent relations from each parse tree.

## 4 Data Sets:

	ENGLISH	GERMAN
TEXT	1,521,998 sentences	808,146 sentences
TOKENS	36,251,168 words	14,311,788 words
TYPES	276,402 words	388,291 words

# IMPLEMENTATION



**1** CORPUS  
CLEANING

**2** PCFG  
PARSING

**3** PS TO DS  
CONVERSION

**4** DATA  
SETS

July 24, 2008

**{ system }**

SEED  
LEXICON

PARSED  
TEXT

RAW  
TEXT

**5** CONTEXT GENERATOR

SYN  
VECTORS

POS  
VECTORS

Lexicon Extraction ~ Ankit

13

# { vector }

- ❖ Seed lexicon obtained from a dictionary, identically spelled words, spelling transformation rules.
- ❖ Context vectors have dimension values (co occurrence of word with seed) normalized on seed frequency.

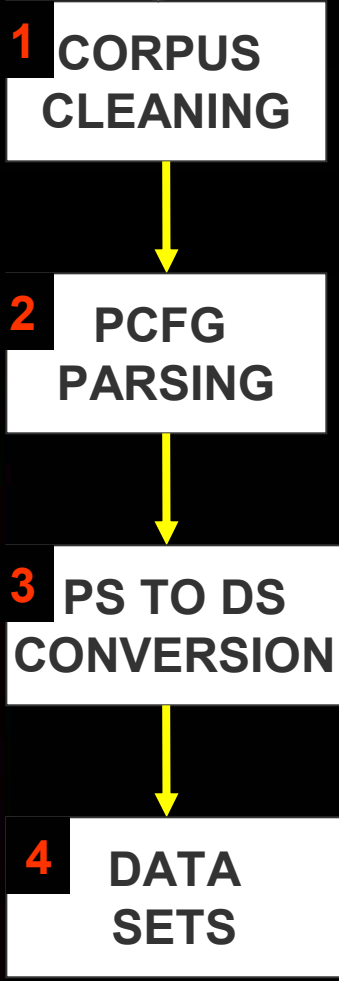
## 5 Context Vectors:

	ENGLISH	GERMAN
DIMENSION	2,376 words	
SEED	2,350 words	2,376 words
UNKNOWN	74,434 words	106,366 words

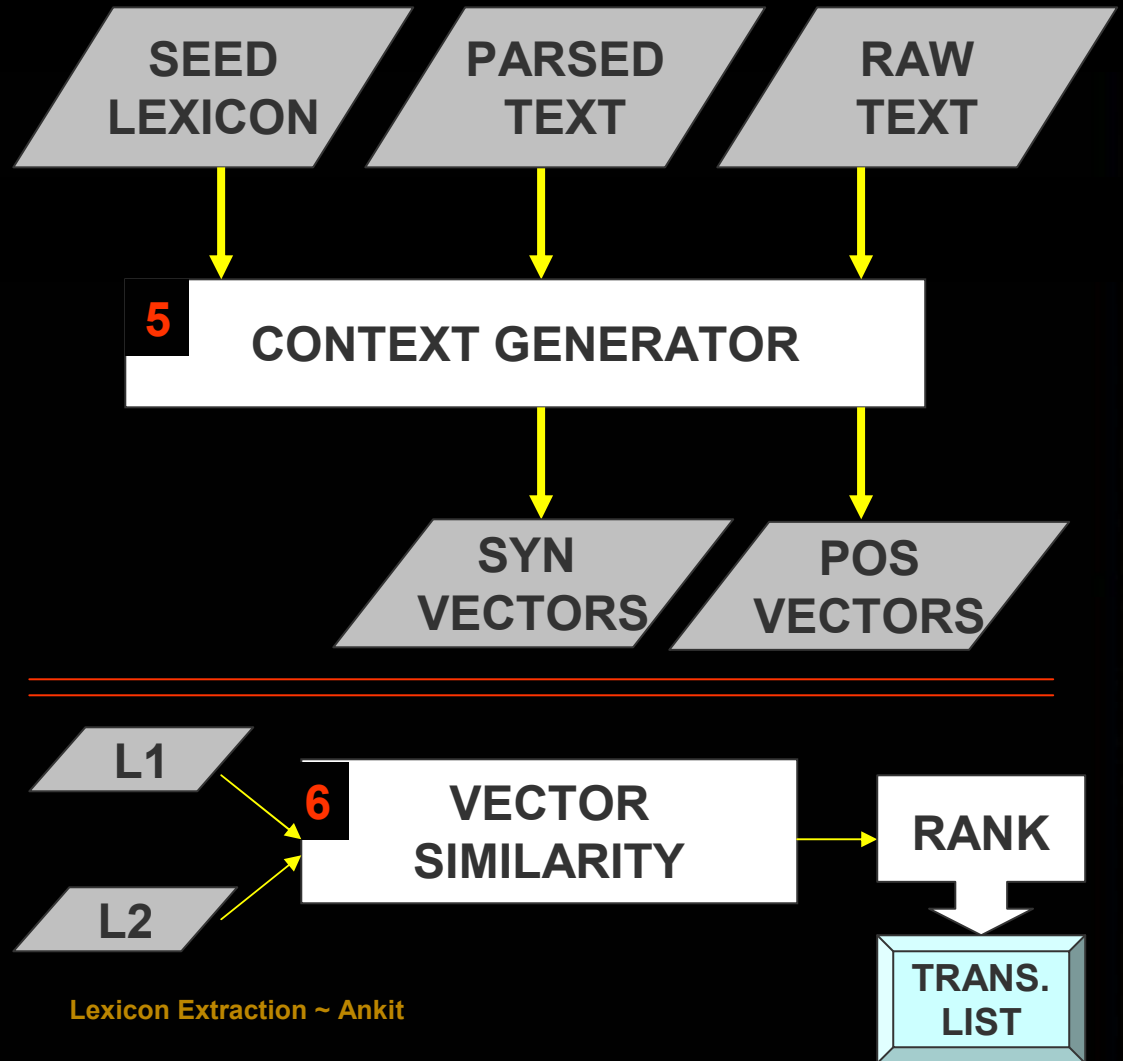
# IMPLEMENTATION



{ system }



July 24, 2008



Lexicon Extraction ~ Ankit

# { evaluate }

- ❖ Vector similarity metrics used are city block [Rapp 1999] and cosine. Translations sorted in descending order of scores.
- ❖ Evaluation data extracted from online bilingual dictionaries (364 translations).

## 6 Ranked Translations Predictor:

	CITY BLOCK	COSINE
POSITIONAL CONTEXT	63 out of 364	148 out of 364
SYNTACTIC CONTEXT	301 out of 364	216 out of 364

## { endnote }

- ❖ Extraction from non parallel corpora useful for compiling lexicon from new domains.
- ❖ Syntactic context helps in focusing the context window, more impact on longer sentences.
- ❖ Non parallel corpora involves more filtering, search heuristics than in parallel.
- ❖ Future directions include using syntactic only on one side, extending coverage through stemming.

**{ thanks }**

