LingFN:

Towards a Framenet for the Linguistics Domain

Per Malm[1] Shafqat Mumtaz Virk[2] Lars Borin[2] Anju Saxena[3]

[1]Department of Scandinavian Languages
Uppsala University, Sweden

[2]Språkbanken
University of Gothenburg, Sweden

[3]Department of Linguistics and Philology
Uppsala University, Sweden

The International FrameNet Workshop 2018, May 2018

Overview

Introduction

- FrameNet: A lexico-semantic resource based on frame-semantics (Fillmore, 1976; Fillmore, 1977; Fillmore, 1982)
- A collection of more than a thousand frames together with the associated frame-elements (FE) and lexical-units (LU)
- Frame-to-Frame and FE-to-FE relations
- Annotated examples and annotated running text

Introduction

- Applications in the areas of information extraction, question answering, coreference resolution, paraphrase extraction, and machine translation (reference in paper)
- Because of its usefulness FrameNet has been developed for a number of languages: Chinese, French, German, Hebrew, Korean, Italian, Japanese, Portuguese, Spanish, and Swedish

Introduction

- Limited coverage of both the network and the annotated data
- A proposed reasonable-effort solution to this problem is to develop domain-specific (sub-language) framenets
- There already exist domain-specific framenets: (1) a framenet to cover medical terminology (2) Kicktionary, a soccer language framenet (3) the Copa 2014 project, covering the domains of soccer, tourism and the World Cup in Brazilian Portuguese, English and Spanish

Introduction

- We report our attempts and initial results of building a domain-specific framenet to cover the concepts and terms used in traditional descriptive linguistic grammars
- Linguistics has developed a rich set of specific terms and concepts (e.g. inflection, agreement, affixation, etc.)
- So we develop a framenet for the linguistic domain

Introduction

- The work is part of another project where attempts are being made to develop methodologies for automatic extraction of linguistic features from descriptive grammars
- We have attempted pattern matching and dependency parsing based approaches (Virk et al. (2017), but we believe a methodology based on the well-established theory of frame semantics is a better option as it offers more flexibility and has proved useful in the area of information extraction in general

The Data

- The Linguistic Survey of India (LSI) (Grierson, 1903–1927)
- It comprises 19 volumes of around 9500 pages in total
- The survey covers 723 linguistic varieties representing major language families and some unclassified languages,

Methodology

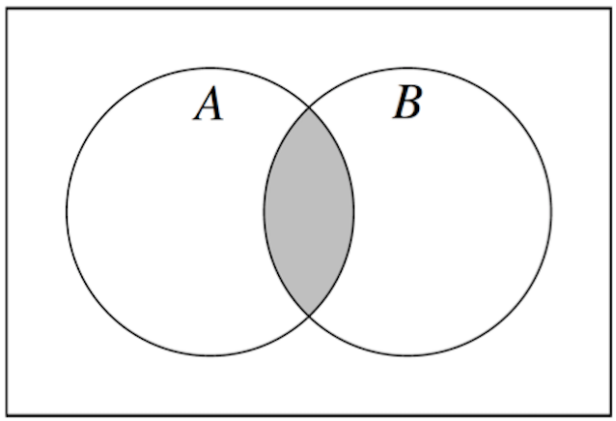
- Method for framenet development?
  - The corpus-driven approach
- Method for dealing with polysemous occurrences of terms?
  - Semiautomatic uniqueness differentiation (SUDi)

Semiautomatic Uniqueness Differentiation

1. collect cases containing the presumed polysemous form
2. sort these intuitively into two text-files
3. process the files using an annotation device that produces xml
4. run the xml-files through *Uneek* (Malm et al., 2018)
5. interpret the result (*proof by contradiction*)

Uneek

- a web tool for comparative analysis of annotated texts
- takes two input files (txt or xml)
- performs set comparison operations, i.e. the *difference* and the *intersection* (Cantor, 1915)



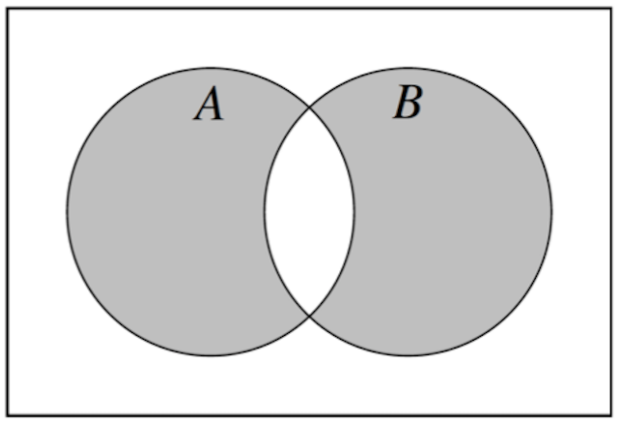


Figure: The intersection $A \bigcap B$ and the differences $A - B$ and $B - A$

Proof by Contradiction

- Take the unique linguistic unit from A and place it in B.
- Does this lead to a 'bad' change in meaning (marked with #)?
  - positive formal support for polysemy
- Does this lead to a 'good' change in meaning?
  - negative formal support for polysemy

A Methodological Example: PLACING Verbs

Is there a difference between the BFN PLACING verbs and Ling PLACING verbs?

(1) The subject is {inserted / placed / put} in the agent case before the verb. (DESK)

(2) The keycard is {inserted / placed / put} in the leftmost slot before the final boss. (DESK)

BFN PLACING frame vs. LING_PLACING: Preliminaries

1. We looked for relevant frames evoked by *insert*, *place*, and *put* in the BFN.
   - The linguistic examples were closely connected to the PLACING frame.
2. We checked if the other LUs in the PLACING frame occurred in the LSI.
   - They did not.

BFN PLACING frame vs. LING_PLACING: Preliminaries

(3) # The subject is { bagged / boxed / caged / jammed / stashed / stuffed / tucked } in the agent case before the verb. (DESK)

SUDi: BFN PLACING frame vs. LING_PLACING

1. All occurrences were collected from the LSI corpus using *Korp* (Borin et al., 2012)
2. The Ling domain cases were manually sorted out from the general ones (1475 Gen domain and 530 Ling domain).
3. The Ling domain cases and the BFN examples, were annotated using the *Stanford parser*: dependencies and POS. (Manning et al., 2014).
4. The Ling cases were compared with the BFN cases using *Uneek* (Malm et al., 2018).
5. Unique occurrences were interpreted, *proof by contradiction*.

Uneek results for POS

Table: Unique POS for Gen domain PLACING verbs

Gen *insert*		Gen *place*		Gen *put*	
PRP$ 'Possessive pronoun'	10	VB 'Verb, base form'	26	JJR 'adjective comparative'	2
WRB 'Wh-adverb'	3	PRP$ 'Possessive pronoun'	13	WP 'Wh- pronoun'	2
–	–	MD 'Modal'	7	JJS 'adjective superlative'	1
–	–	WRB 'Wh-adverb'	7	–	–
–	–	JJR 'adjective, comparative'	3	–	–
–	–	JJS 'adjective, superlative'	1	–	–

Interpreting Uneek results for POS: Possessive Pronouns

(4) a. . . . the anaesthetist had to insert a tube in *his* mouth . . . (BFN)

b. Eadmer inserted them . . . into *his* Historia Novorum. (BFN)

(5) a. The verb is inflected; a prefix is put at $\left\{\begin{array}{r} * \text{ his} \\ * \text{ her} \\ \text{its} \end{array}\right\}$ base. (DESK)

Interpreting Uneek results for POS: Verbs

(6) a. Adjectives $\left\{\begin{array}{l}\#\ \textit{should have been}\\ \#\ \textit{might have been}\\ \#\ \textit{will have to be}\\ \#\ \textit{would be}\\ \textit{may be}\\ \textit{are}\end{array}\right\}$ placed after the noun. (DESK)

b. The adjective $\left\{\begin{array}{l}\#\ \textit{had been placed}\\ \#\ \textit{has been placed}\\ \#\ \textit{is being placed}\\ \#\ \textit{was placed}\\ \textit{is placed}\end{array}\right\}$ after the noun. (DESK)

Uneek results for words

Table: Top ten unique PLACING words in the Gen domain

Gen *insert*		Gen *place*		Gen *put*	
into	9	place	18	his	15
his	6	on	14	she	15
he	5	he	7	her	11
through	5	them	7	against	10
under	5	has	6	he	7
text	4	under	6	through	7
's	3	against	5	my	6
computer	3	from	5	's	5
left	3	her	5	arm	5
new	3	should	5	said	5

Interpreting Uneek results for words

(7) a. Adverbs are put $\left\{\begin{array}{l}\text{after}\\ \text{before}\\ \text{between}\\ \text{at the end of}\\ \text{at the beginning of}\end{array}\right\}$ verbs. (DESK)

Interpreting Uneek results for words

(8) a. Adverbs are put $\left\{\begin{array}{l}\text{after}\\ \text{before}\\ \text{between}\\ \text{at the end of}\\ \text{at the beginning of}\end{array}\right\}$ verbs. (DESK)

b. # Adverbs are put $\left\{\begin{array}{l}\text{on}\\ \text{into}\\ \text{under}\\ \text{against}\\ \text{through}\end{array}\right\}$ verbs. (DESK)

Interpreting Uneek results for words

(9) a. Adverbs are put $\left\{\begin{array}{l}\text{after}\\ \text{before}\\ \text{between}\\ \text{at the end of}\\ \text{at the beginning of}\end{array}\right\}$ verbs. (DESK)

b. # Adverbs are put $\left\{\begin{array}{l}\text{on}\\ \text{into}\\ \text{under}\\ \text{against}\\ \text{through}\end{array}\right\}$ verbs. (DESK)

c. But, uncomfortable chairs are put $\left\{\begin{array}{l}\text{on the floor}\\ \text{under the table}\\ \text{against something brittle}\\ \text{into corrosive compounds}\\ \text{through my secret machine}\end{array}\right\}$. (DESK)

Uneek results for dependencies

Table: Unique dependencies for the Gen and Ling domain

General domain *insert*		Linguistic domain *insert*	
NMOD:NPMOD 'NP as adverbial modifier'	3	NEG 'negation'	12
–	–	PARATAXIS 'discourse-like coordination'	10
–	–	MWE ' multi-word expression'	2
–	–	DET:PREDET 'predeterminer'	1
–	–	DISCOURSE 'discourse particles'	1
General domain *place*		**Linguistic domain *place***	
AUX 'non-main verb, auxiliary'	23	COP 'copula'	4
NMOD:POSS 'possessive nominal modifier'	17	–	–
NMOD:NPMOD 'noun phrase as adverbial modifier'	3	–	–
MWE 'multiword expression'	2	–	–
General domain *put*		**Linguistic domain *put***	
NMOD:TMOD 'temporal modifier'	1	NSUBJPASS 'passive nominal subject'	165
–	–	CC:PRECONJ 'preconjunct e.g. *both*'	2
–	–	EXPL 'expletive, e.g. *there*-Cx'	1
–	–	NMOD:NPMOD 'NP as adverbial modifier'	1

Interpreting Uneek results for dependencies

(10) a. # Adjectives are placed after the noun $\underbrace{\textit{by the speaker}}_{\text{AGENT}}$. (DESK)

b. # $\underbrace{\textit{The speaker}}_{\text{AGENT}}$ places the adjective after the noun. (DESK)

Interpreting Uneek results for dependencies

(11) a. # Adjectives are placed after the noun $\underbrace{\textit{by the speaker}}_{\text{AGENT}}$. (DESK)

b. # $\underbrace{\textit{The speaker}}_{\text{AGENT}}$ places the adjective after the noun. (DESK)

c. $\underbrace{\underbrace{\textit{all}}_{\text{PREDET}} \textit{ these verbs}}_{\text{CAUSE}}$ insert *m* in the [. . .] singular [. . .]. (LSI)

Result in Sum

- Positive formal support for polysemous PLACING words:
  - no possessive pronouns with animate reference
  - tense restrictions
  - modality restrictions
  - prepositional restrictions
  - voice restrictions[1]

[1] Exceptions: anthropomorphism, CAUSES in *insert.v*

Result in Sum

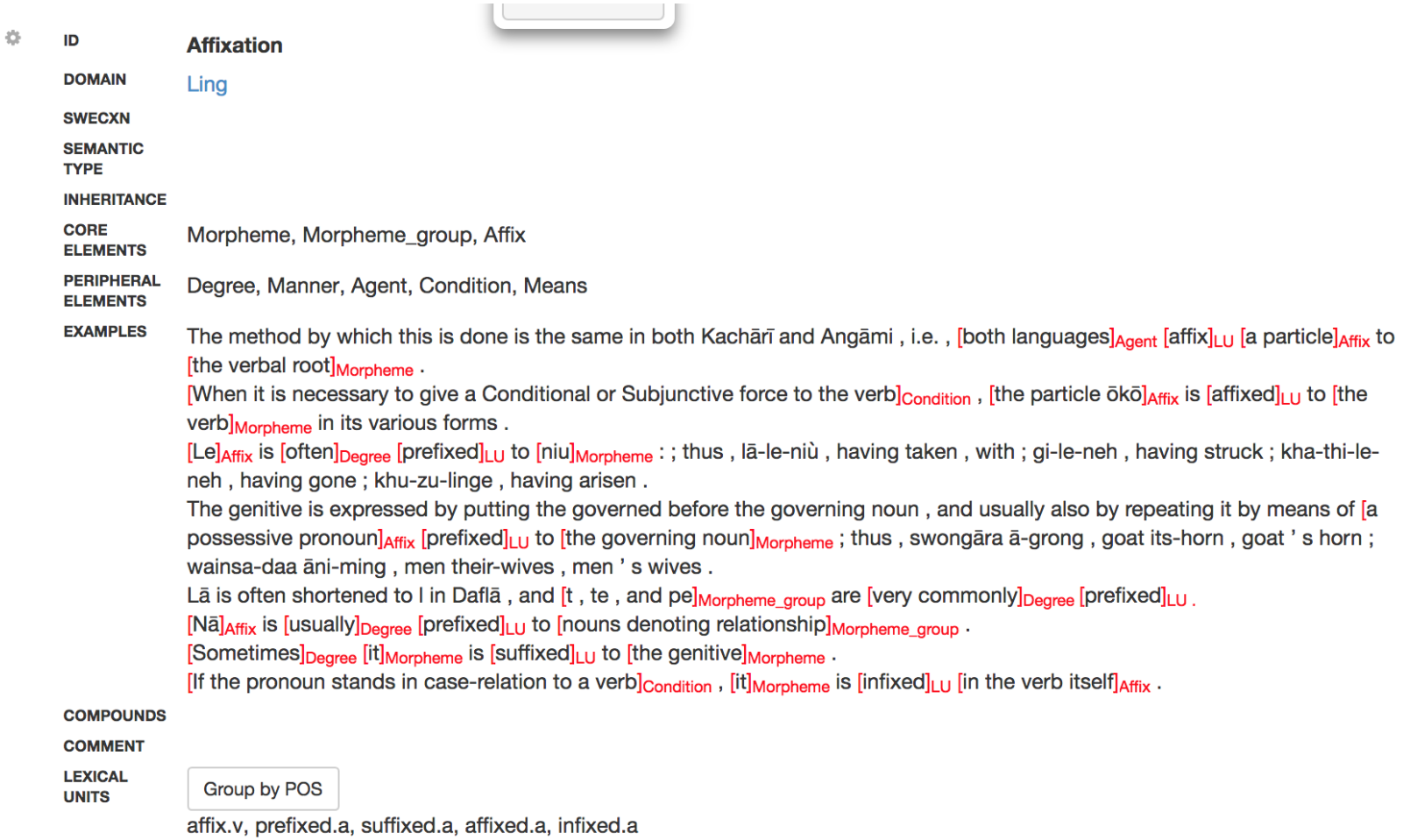
- Positive formal support for polysemous PLACING words:
  - no possessive pronouns with animate reference
  - tense restrictions
  - modality restrictions
  - prepositional restrictions
  - voice restrictions[1]
- Bulletproof analysis? No. For the LSI perhaps and for our purposes of finding rule-like statements, but probably not for a larger collection of corpora.

[1] Exceptions: anthropomorphism, CAUSES in *insert.v*

Developed Linguistics Domain Frames

Types	Number of types
Frames	12
Core and non-core frame elements	74
Annotated example sentences	156
Lexical units	106

Developed Linguistics Domain Frames



ID	**Affixation**
DOMAIN	Ling
SWECXN	
SEMANTIC TYPE	
INHERITANCE	
CORE ELEMENTS	Morpheme, Morpheme_group, Affix
PERIPHERAL ELEMENTS	Degree, Manner, Agent, Condition, Means
EXAMPLES	The method by which this is done is the same in both Kachārī and Angāmi , i.e. , [both languages]$_{Agent}$ [affix]$_{LU}$ [a particle]$_{Affix}$ to [the verbal root]$_{Morpheme}$. [When it is necessary to give a Conditional or Subjunctive force to the verb]$_{Condition}$, [the particle ōkō]$_{Affix}$ is [affixed]$_{LU}$ to [the verb]$_{Morpheme}$ in its various forms . [Le]$_{Affix}$ is [often]$_{Degree}$ [prefixed]$_{LU}$ to [niu]$_{Morpheme}$: ; thus , lā-le-niù , having taken , with ; gi-le-neh , having struck ; kha-thi-le-neh , having gone ; khu-zu-linge , having arisen . The genitive is expressed by putting the governed before the governing noun , and usually also by repeating it by means of [a possessive pronoun]$_{Affix}$ [prefixed]$_{LU}$ to [the governing noun]$_{Morpheme}$; thus , swongāra ā-grong , goat its-horn , goat ' s horn ; wainsa-daa āni-ming , men their-wives , men ' s wives . Lā is often shortened to l in Daflā , and [t , te , and pe]$_{Morpheme\_group}$ are [very commonly]$_{Degree}$ [prefixed]$_{LU}$. [Nā]$_{Affix}$ is [usually]$_{Degree}$ [prefixed]$_{LU}$ to [nouns denoting relationship]$_{Morpheme\_group}$. [Sometimes]$_{Degree}$ [it]$_{Morpheme}$ is [suffixed]$_{LU}$ to [the genitive]$_{Morpheme}$. [If the pronoun stands in case-relation to a verb]$_{Condition}$, [it]$_{Morpheme}$ is [infixed]$_{LU}$ [in the verb itself]$_{Affix}$.
COMPOUNDS	
COMMENT	
LEXICAL UNITS	Group by POS affix.v, prefixed.a, suffixed.a, affixed.a, infixed.a

Figure: An example frame from the linguistic domain

Conclusions and Future Work

- In sum, we have:
  - motivated developing a framenet for the linguistic domain
  - presented a methodology for judging the polysemous nature of lemmas in a given corpus, and for identifying domain-specific occurrences
  - reported on our progress so far.
- In the future we wish to:
  1. extend the set of domain specific frames
  2. annotate other descriptive grammars
  3. use the annotations to train a parser for purposes of automated extraction of linguistic features.

Thank you!

References I

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012* (pp. 474–478). Istanbul: ELRA. Retrieved from `http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf`

Cantor, G. (1915). *Contributions to the Founding of the Theory of Transfinite Numbers* (No. 1). Open Court Publishing Company.

Grierson, G. A. (1903–1927). *A linguistic survey of India* (Vols. I–XI). Calcutta: Government of India, Central Publication Branch.

Malm, P., Ahlberg, M., & Rosén, D. (2018). Uneek: a web tool for comparative analysis of annotated texts. In *Proceedings of the IFNW 2018 Workshop on Multilingual FrameNets and Constructicons at LREC 2018.* Miyazaki: ELRA.

References II

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60).