Building Valency Frames Automatically

Ondřej Bojar
obo@cuni.cz

June 8, 2006

Outline

- An extremely brief introduction to valency and VALLEX.
- VALLEX coverage: motivation for building frames automatically.
- Evaluation metrics for frame generation.
- Approaches to frame generation.
- Summary and open issues.

Valency

- Valency is the ability to bind/require modifications of a particular type.
- Specific valency patterns of words are captured in a dictionary.
- Semantic clustering of verbs correlates with clustering of syntactic patterns.

VALLEX Structure

Word entry

Frame entry

odpovídat (imperfective)

$\boxed{1}$ odpovídat$_{1}$ ∼ odvětit [answer; respond]

- frame: $\mathrm{ACT}^{obl}{}_{1}\ \mathrm{ADDR}^{obl}{}_{3}\ \mathrm{PAT}^{opt}{}_{na+4,4}\ \mathrm{EFF}^{obl}{}_{4,aby,a\check{t},zda,\check{z}e}\ \mathrm{MANN}^{typ}$
- example: *odpovídal mu na jeho dotaz pravdu / že . . .* [he responded to his question truthfully / that . . .]
- asp.counterpart: odpovědět$_{1}$ pf.
- class: communication

$\boxed{2}$ odpovídat$_{2}$ ∼ reagovat [react]

- frame: $\mathrm{ACT}^{obl}{}_{1}\ \mathrm{PAT}^{obl}{}_{na+4}\ \mathrm{MEANS}^{typ}{}_{7}$
- example: *pokožka odpovídala na včelí bodnutí zarudnutím* [the skin reacted to a bee sting by turning red]
- asp.counterpart: odpovědět$_{2}$ pf.

. . .

odpovídat se (imperfective)

$\boxed{1}$ odpovídat se$_{1}$ ∼ být zodpovědný [be responsible]

- frame: $\mathrm{ACT}^{obl}{}_{1}\mathrm{ADDR}^{obl}{}_{3}\mathrm{PAT}^{obl}{}_{z+2}$
- example: *odpovídá se ze ztrát* [he answers for the losses]

An abbreviated example for the base lemma “odpovídat”.
Key components: Frames, functors, obligatoriness, morphemic form(s).

VALLEX Coverage

	Occ.	[%]	Verb lemmas	[%]
		VALLEX 1.0		
Covered	8.0M	53.7	1,064	3.6
Not covered but frequent	4.1M	27.9	20	0.1
Not covered, infrequent	2.7M	18.3	28,385	96.3
Total	14.8M	100.0	29,469	100.0
		VALLEX 1.5		
Covered	8.0M	65.6	1,802	6.1
Not covered but frequent	3.5M	23.4	4	0.0
Not covered, infrequent	1.6M	10.9	27,663	93.9
Total	14.8M	100.0	29,469	100.0

⇒ attempt at learning frames for unseen verbs, automatically.

Evaluation Metrics for Frame Generation (FG)

If a system suggests frames for a verb, how do we tell the system was correct?

- Frame precision/recall (Korhonen [2002]).
- Slot precision/recall (Sarkar and Zeman [2000]).

$$Precision = \frac{correctly\ suggested\ frames/slots}{frames/slots\ suggested}$$

$$Recall = \frac{frames/slots\ suggested}{frames/slots\ needed}$$

- Frame Edit Distance and Entry Similarity (Benešová and Bojar [2006]).

Frame Edit Distance (FED)

FED = the number of edit operations (insert, delete, replace) necessary to convert a hypothesized frame to a correct frame:

- currently equal costs of all basic edit operations (fixing the obligatoriness flag, adding or removing allowed morphemic forms).
- to change the functor, one pays for complete destruction of the wrong slot and complete construction of the correct slot.
- we consider to charge more for slot destruction that for slot construction, because we generally prefer frames that possibly miss some information to frames that contain incorrect information.

Verb Entry Similarity (ES)

Given a verb lemma, the set of its VALLEX entries and a set of entries produced by an automatic frame suggestion method, we define ENTRY SIMILARITY or EXPECTED SAVING (ES):

$$ES(G,H) := 1 - \frac{\min FED(G,H)}{FED(G,\emptyset)+FED(H,\emptyset)}$$

G denotes the set golden verb entries of this base lemma
H denotes the hypothesized entries
$\emptyset$ stands for a blank verb entry

Not suggesting anything has ES of 0% and suggesting the golden frames exactly has ES of 100%.

. . . ES estimates how much of lexicographic labour was saved.

Baseline: ACT(1): $ES \sim 27\%$; ACT(1) PAT(4): $ES \sim 38\%$

Overview of Approaches to Frame Generation

- Treat frames as opaque symbols and:
  - Reuse word-frame disambiguation (Bojar et al. [2005])
    Originally WFD was restricted to *known* verbs, relax this requirement.
  - Use similarity of verb occurrences to suggest known frames to new verbs.
  - Convert frames to prototypical patterns to search for in a corpus.

- Decompose frames into parts (reflexivity, slots, functors, morphemic forms, oblig.) and use corpus evidence to suggest frame parts to occurrences of new verbs.

Finally, collect/clean up the set of frames seen with a particular verb
$\approx$ cluster verb occurrences into groups with similar/same frame.

Deep Syntactic Distance (DSD)

Given two verb occurrences v_1, v_2, DSD estimates how different the verbs' frames are, based typically on the verbs' surface modifications ${m_1}^1 \ldots {m_1}^i$ and ${m_2}^1 \ldots {m_2}^j$.

DSD captures, how difficult is to assume that ${m_1}^x$ and ${m_2}^y$ both express the same slot in the frame, i.e. both share the same functor f. The pairing is chosen so that the total cost is minimum for all modifications x and y.

$$\begin{aligned} DSD(v_1, v_2) := \quad & \min_{p \textit{ pairing of} \{m_1\} \textit{ and } \{m_2\}} \\ & \sum_{(x,y) \in p} \min_{f \in Functors} cost({m_1}^x, f) cost({m_2}^y, f) \end{aligned}$$

The $cost(m, f)$ is estimated based on functor-form co-occurrence statistics in PDT. (E.g. $cost(nominative, ACT) < cost(dative, ACT)$)

DSD Sometimes Mismatches Human Annotation

The DSD estimates very near distance (the sets of sons are nearly equal) of the following three occurences of *ležet (lie)*, but the frames are different.

Sentence ID	Frame ID	Text
[ln94203-1-p3s2]	v-w1699f1	<u>Leží</u> **v** jedné z biologicky nejproduktivnějších oblastí světového oceánu. . .
[lnd94103-087-p1s74]	v-w1699f2	Často jsme ho našli , jak <u>leží</u> **zablácený v** posteli .
[mf930713-162-p2s5]	v-w1699f1	V tomto případě <u>ležel</u> **mrtvý ve** svém domě. . .

v-w1699f1: ACT(.1) LOC(*): ležet na dně oceánu, l. jižně od Prahy

v-w1699f2: ACT(.1): nemůže ležet, protože ho bolí záda

⇒ DSD can be used to search for suspicious annotations.

⇒ DSD should be extended to capture the semantic (e.g. WordNet) distance between the sons.

Frames as Prototypical Patterns (ProtPat)

Benešová and Bojar [2006] describe a particular instance of this approach:

- Verbs of communication usually allow for the frame ACT+ADDR+PAT (speaker, addresse and the content conveyed).
- Using the morphemic realizations listed in the dictionary, the frame can be converted to a corpus pattern:

$$Verb + Noun/Pronoun[case : 2|3|4] + SubordinateClause$$

- Verbs appearing in this pattern tend to belong to the communication class (and allow for this particular pattern) $\Rightarrow$ slight ES improvement.

A similar approach can be followed for all frames.

Decomposing Frames (Decomp)

- Objects: verb occurrences in PDT 2.0 (PDT-VALLEX frames known).
- Input features: morphological and surface syntactic info about the verb (similar to WFD by Jiří Semecký).
- Output features: features about the frame assigned to the verb occ:
  - has_ftor(ACT) . . . yes/no
  - slot_type(PAT, oblig) . . . yes/no
- Use a machine learning technique to predict each of the output features given input features.
- Baseline accuracy: predict the most common value, e.g. has_ftor(ACT): yes

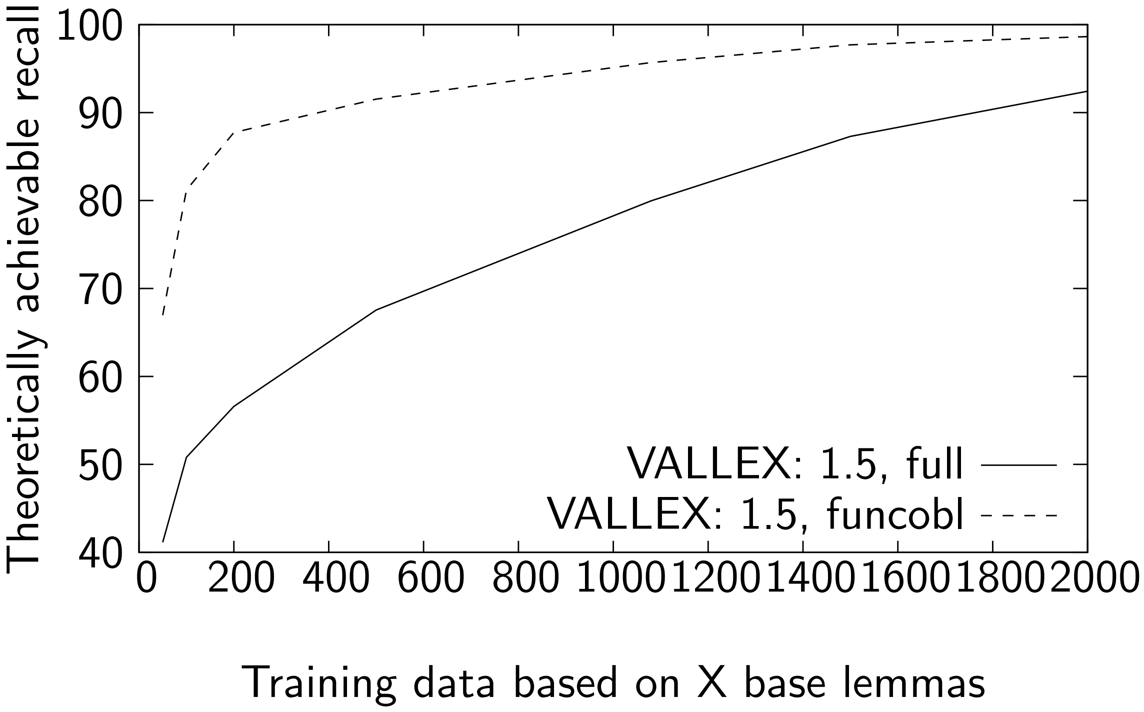
Problem with Decomp: Zipfian Distribution

- Many of the output features are too easy (baseline too high), because one of the values is extremely dominant.

Baseline Acc	Achieved Acc	Improvement	Target (Output) Feature
84.6	94.2	9.6	has_ftor(ADDR)
79.8	86.7	6.8	has_ftor(PAT)
89.4	95.3	6.0	has_ftor(EFF)
90.8	96.0	5.1	has_ftor(ORIG)
95.7	97.2	1.5	has_ftor(DIR3)
98.4	99.1	0.7	has_ftor(DIR1)
98.2	98.4	0.2	has_ftor(LOC)
99.4	99.6	0.2	has_ftor(EXT)
100.0	100.0	0.0	has_ftor(TFRWH)
100.0	100.0	0.0	has_ftor(ACMP)

$\Rightarrow$ PDT alone insufficient to guess presence of most functors in verb frames.

Achievable Recall when Suggesting Whole Frames



⇒ only functors and obligatoriness can be considered if frames are taken as indivisible wholes.

Summary and Open Issues

- Still hoping that missing (low-frequency) verbs are easier.
- A novel metric FED proposed for estimating the lexicographic labour saved.
- PDT seems insufficient even for learning frame parts (Decomp)
  $\Rightarrow$ adding non-annotated data is a must.
- Three methods assign frames to verb occurrences: WFD, DSD, ProtPat.
- The repertoire of frames (ftors+oblig) seems to be nearly closed.
- Open issues:
  - Clustering the set of verb occurrences into groups with similar frames.
  - Combining the frames in each cluster into a common representative.
  - Additional goal: Attempt at automatic estimation of cluster count.

References

Václava Benešová and Ondřej Bojar. Czech Verbs of Communication and the Extraction of their Frames. In *Text, Speech and Dialogue: 9th International Conference, TSD 2006*, volume LNAI 3658. Springer Verlag, September 2006. ISBN 3-540-28789-2. (in print).

Ondřej Bojar, Jiří Semecký, and Václava Benešová. VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics*, 83:5–17, 2005. ISSN 0032-6585.

Anna Korhonen. Subcategorization Acquisition. Technical Report UCAM-CL-TR-530, University of Cambridge, Computer Laboratory, Cambridge, UK, February 2002.

Anoop Sarkar and Daniel Zeman. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics (Coling 2000)*, Saarbrücken, Germany, 2000. Universität des Saarlandes.