Unsupervised Frame Induction Through PCFGs

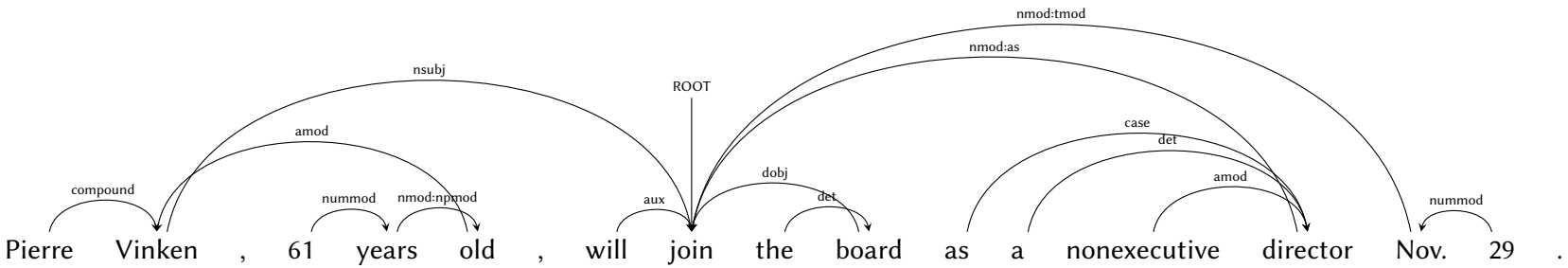
Laura Kallmeyer & Behrang QasemiZadeh

Heinrich-Heine-Universität Düsseldorf

CL Colloquium

Summer 2017

Introduction



join.f1 / ew1777f1 / 22.1-2-1 / Cause-to-amalgamate

Vinken	ACT-arg / Agent / ARG0
board	PAT-arg / Patient / ARG1
director	COMPL / _ / ARGM-PRD
Nov.	TWHEN / _/ ARGM-TMP

Introduction: Frame Induction Task

- Identifying things which will be presented as frames;
- Finding informative sub-structures: roles, slots, etc.;
- Labelling/naming the identified frames;
- Labelling/naming the identified substructures.

Introduction: Frame Induction Task

- **Identifying things which will be presented as frames;**
- **Finding informative sub-structures: roles, slots, etc.;**
- Labelling/naming the identified frames;
- Labelling/naming the identified substructures.

Introduction: **Supervised** Frame Induction

In the supervised scenario, a machine learning method is provided with manually annotated data:

1	Pierre	Pierre	NNP	-	-	_	NE	_	_	_	_	_
2	Vinken	vinken	NNP	-	+	_	_	_	_	ACT-arg	_	_
3	,	,	,	-	-	_	_	_	_	_	_	_
4	61	61	CD	-	-	_	_	RSTR	_	_	_	_
5	years	year	NNS	-	+	_	_	_	EXT	_	_	_
6	old	old	JJ	-	+	_	DESCR	_	_	_	_	_
7	,	,	,	-	-	_	_	_	_	_	_	_
8	will	will	MD	-	-	_	_	_	_	_	_	_
9	join	join	VB	+	+	ev-w1777f1	_	_	_	_	_	_
10	the	the	DT	-	-	_	_	_	_	_	_	_
11	board	board	NN	-	-	_	_	_	_	PAT-arg	_	_
12	as	as	IN	-	-	_	_	_	_	_	_	_
13	a	a	DT	-	-	_	_	_	_	_	_	_
14	nonexecutive	nonexecutive	JJ	-	-	_	_	_	_	_	RSTR	_
15	director	director	NN	-	+	_	_	_	_	COMPL	_	_
16	Nov.	nov.	NNP	-	+	_	_	_	_	TWHEN	_	_
17	29	29	CD	-	-	_	_	_	_	_	_	RSTR
18	.	.	.	-	-	_	_	_	_	_	_	_

Introduction: **Unsupervised** Frame Induction

- In unsupervised methods, there is no manually annotated data to train a model.
- One way to realize an unsupervised method is to assume that "frames and semantic roles" are latent factors of a probabilistic model that describes their distribution in a corpus.
- We then use the expectation maximization algorithm (EM) to estimate parameters of the supposed probabilistic model with incomplete data.

Introduction: **Unsupervised** Frame Induction

Assumptions:

- The focus is only on frames that are evoked by verbs (i.e., any verb in the corpus).
- We know the maximum number of frames and semantic roles and their structural relationship.
- In particular, inspired by Cheung et al. (2013) we assume that:

$$pmf(f_1, \ldots, f_n, s_1 \ldots s_k, d_1, \ldots, d_m; \mathcal{C}, \theta) \approx$$

$$P_\theta(f_1 = v_1, \ldots, f_n = v_n, s_1 \approx (w_1/dep_1), \ldots, s_k \approx (w_k/dep_k), d_1 = dep_1, \ldots, d_m = dep_m) \approx$$

$$P(f) \cdot P(v|f) \cdot \prod_{i=1}^{k} P(s_i|f, d_i) \cdot \prod_{i=1}^{k} P(w_i|s_i)$$

From the Probabilistic Model to the PCFG

Starting point: generative model for a specific frame f with head v, semantic roles $s_1, \ldots, s_k$ filled by words $w_1, \ldots, w_k$ and corresponding syntactic dependencies $d_1, \ldots, d_k$ (note that the dependencies are taken to be generated as well):

Probabilistic Model

$$P_\theta(f, v, s_1 \ldots s_k, w_1 \ldots w_k, d_1 \ldots d_k) \approx$$

$$P(f) \cdot P(v|f) \cdot \prod_{i=1}^{k} P(d_i|f) \cdot \prod_{i=1}^{k} P(s_i|f, d_i) \cdot \prod_{i=1}^{k} P(w_i|s_i)$$

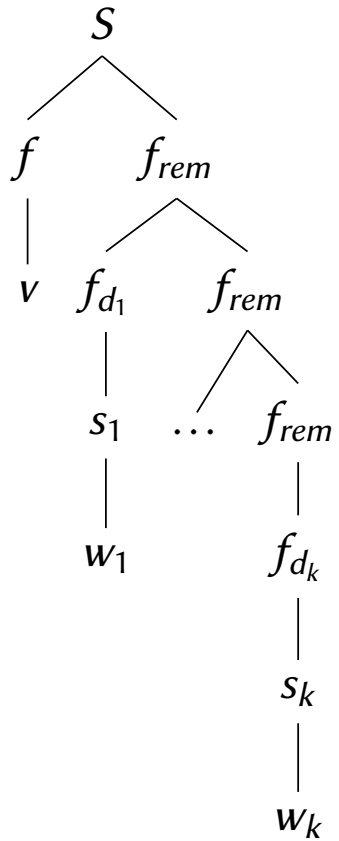
This model describes how to generate a combination of a frame structure with a specific head and specific argument fillers displaying specific syntactic dependencies between head and arguments.

From the Probabilistic Model to the PCFG

- Our proposal: translate the conditional probabilities in this model into PCFG parameters and estimate them using for instance EM.
- In a PCFG, the probability of a rule $A \to \gamma$ is the probability of γ given A, i.e., $P(A \to \gamma) = P(\gamma|A)$.
- Translate $P(w|s)$ into $P(s \to w)$, $P(d|f)$ into $P(f \to f_d f_{rem})$ $P(s|f, d)$ into $P(f_d \to s)$, $P(v|f)$ into $P(f \to v)$ and finally $P(f)$ into $P(S \to f\ f_{rem})$ where S is the CFG start symbol.

From the Probabilistic Model to the PCFG

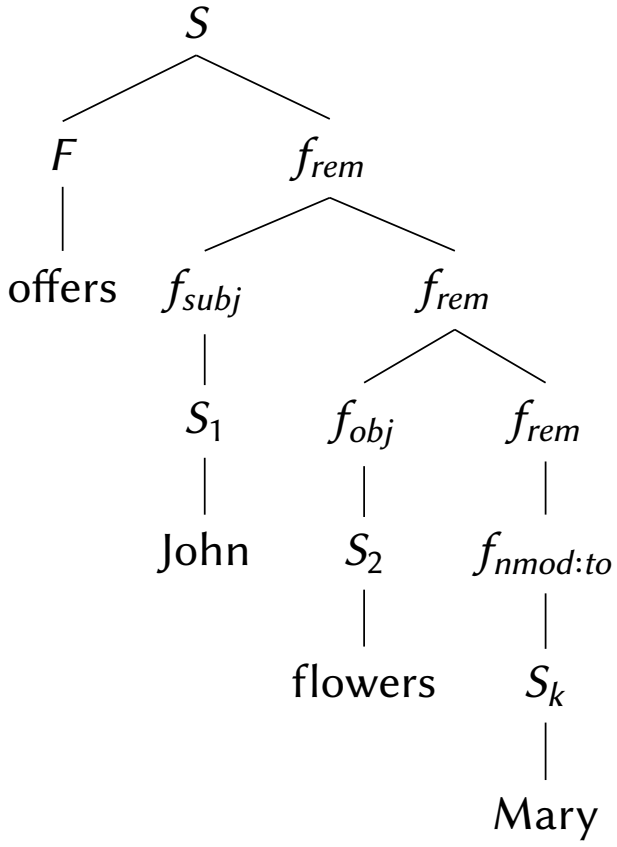
$$P(f) \cdot P(v|f) \cdot \prod_{i=1}^{k} P(d_i|f) \cdot \prod_{i=1}^{k} P(s_i|f, d_i) \cdot \prod_{i=1}^{k} P(w_i|s_i)$$



From the Probabilistic Model to the PCFG

(1) John$_{subj}$ offers flowers$_{dobj}$ to Mary$_{nmod:to}$

Assuming that (1) has frame F with semantic roles S_1, S_2, S_3, this is encoded by the following PCFG tree:



A Short Note on Dependency Parses

- We use dependency parses in the Enhanced Universal Dependencies format (Schuster and Manning, 2016).
- This formalism links "content words" through dependency relations and attaches function words to the words that they modify.
  - E.g., in "I was loved by you", 'love' and 'you' are linked to each other through a *nmod* relation, and 'you' and 'by' through a *case* relation.
- The enhanced version augments dependency relations between content words by adding information distilled from their relations to function words:
  - E.g., in the above, the relation between 'love' and 'you' is *nmod:agent*.
- None of these formalisms distinguishes between (obligatory) arguments and (non-obligatory) adjuncts.

From the Probabilistic Model to the PCFG

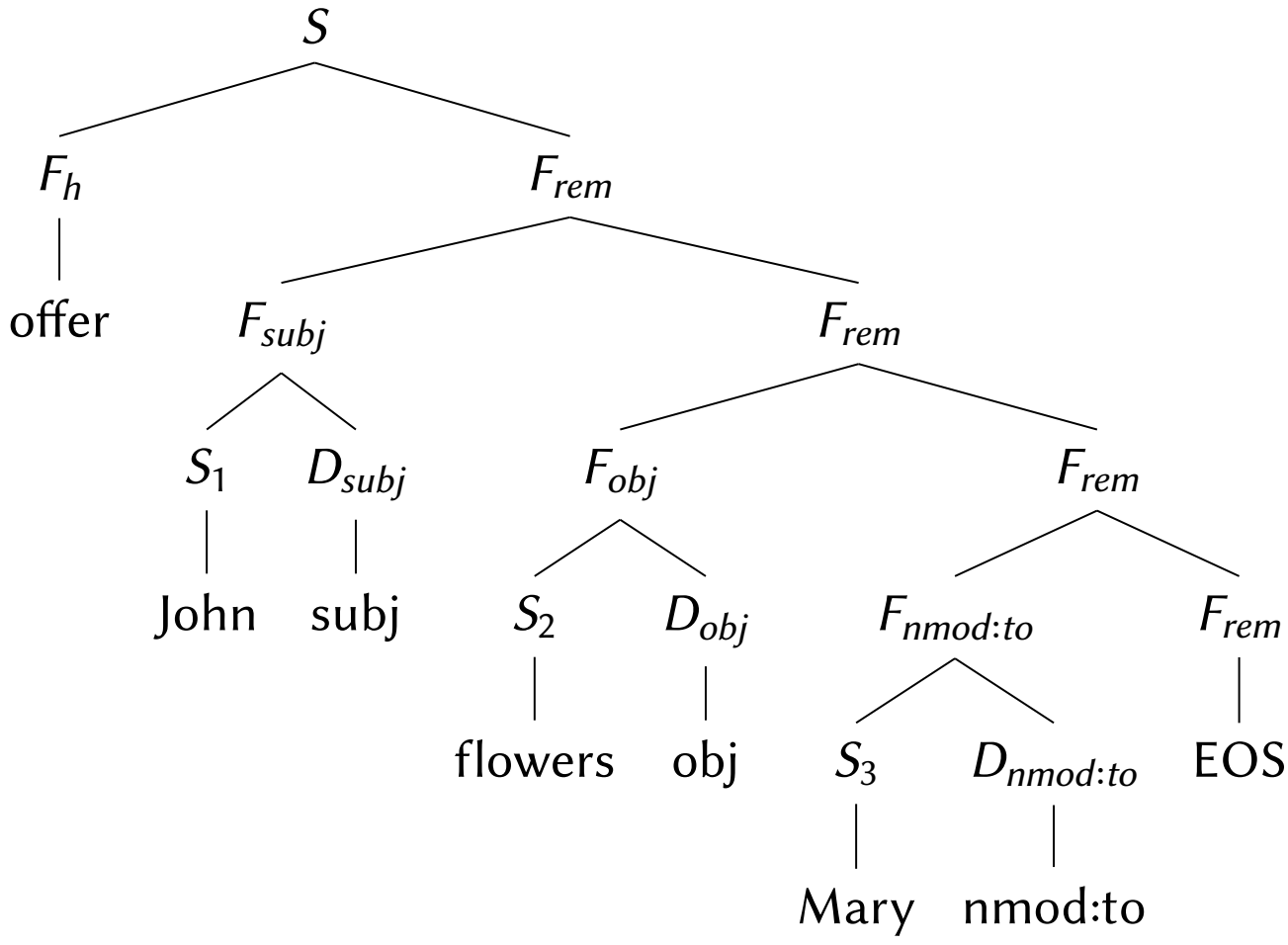
- The dependencies are assumed to be known when doing parsing to frames, i.e., they should also be part of the input string.
- Furthermore, we assume the string to be ordered such that the head is followed by the arguments:

  (2) offer John subj flowers dobj Mary nmod:to EOS

- We want binarized tree structures with preterminals in order to facilitate parameter estimation and PCFG parsing.

From the Probabilistic Model to the PCFG

Therefore, the above structure is transformed to



From the Probabilistic Model to the PCFG

The PCFG contains the following rules (given fixed alphabets $\mathcal{F}, \mathcal{S}, D, T_v, T_n$ of frames, semantic roles, dependencies, verbal heads and role fillers resp. and a start symbol S):

- $S \to f_h\, f_{rem}$ and $f_{rem} \to EOS$ for all $f \in \mathcal{F}$;
- $f_{rem} \to f_d\, f_{rem}$ for all $f \in \mathcal{F}$ and $d \in D$;
- $f_d \to s\, D_d$ for all $f \in \mathcal{F}$, all $s \in \mathcal{S}$ and and $d \in D$;
- $f_h \to v$ for all $f \in \mathcal{F}$ and all $v \in T_v$;
- $s \to n$ for all $s \in \mathcal{S}$ and all $n \in T_n$;
- $D_d \to d$ for all $d \in D$.

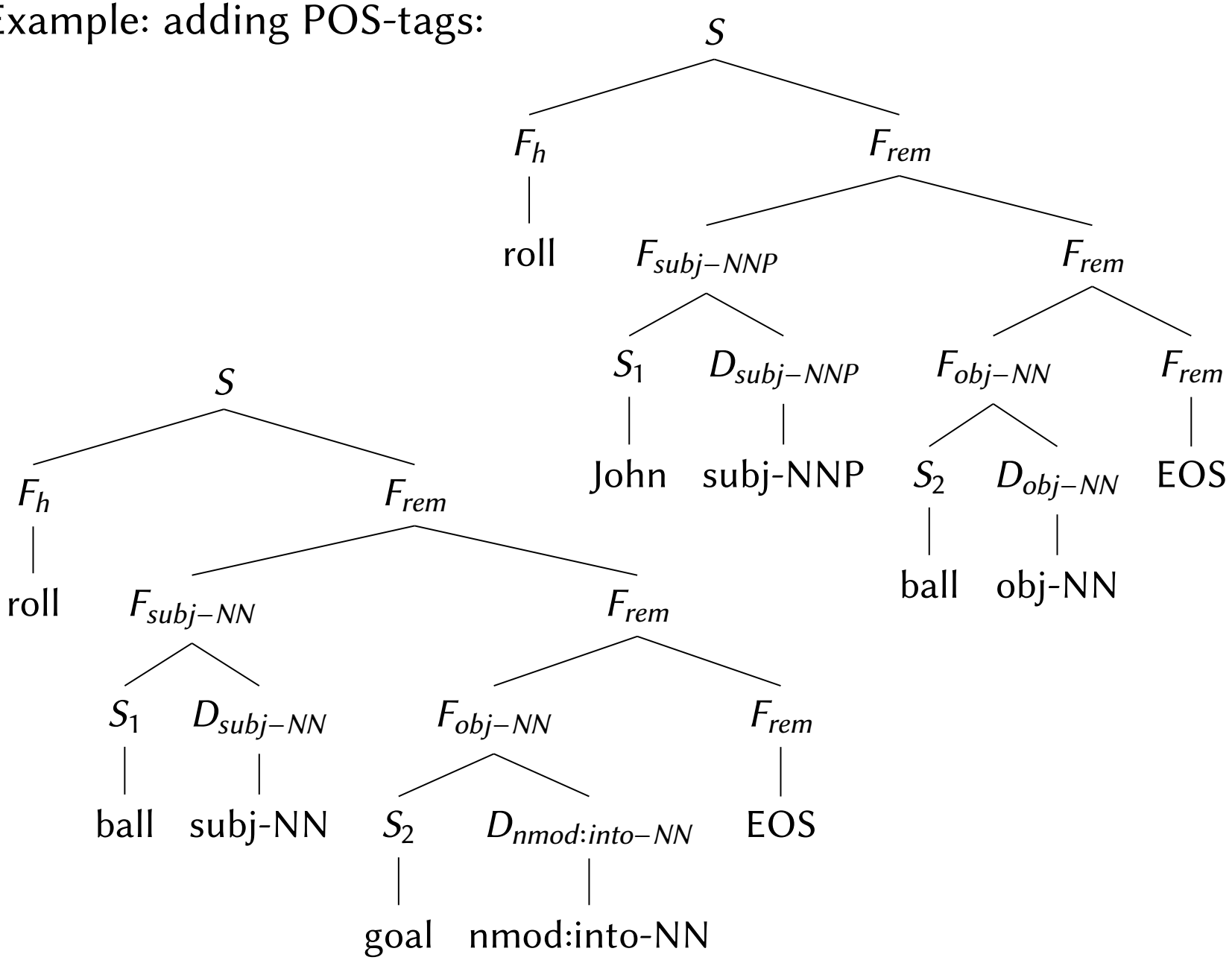
From the Probabilistic Model to the PCFG

Advantages of such a PCFG encoding:

- The frame induction task can be boiled down to a PCFG parameter estimation task.
- Split-and-merge techniques from PCFG induction can be used for frame/role refinement.
- One can easily change the model by just changing the underlying grammar. The parameter estimation stays the same.
- Possible extensions:
  - add more information on the level of the string (e.g., POS tags, Brown clusters, ...)
  - add more dependencies between the non-terminal nodes by lexicalizing or by adding vertical/horizontal context

From the Probabilistic Model to the PCFG

Example: adding POS-tags:



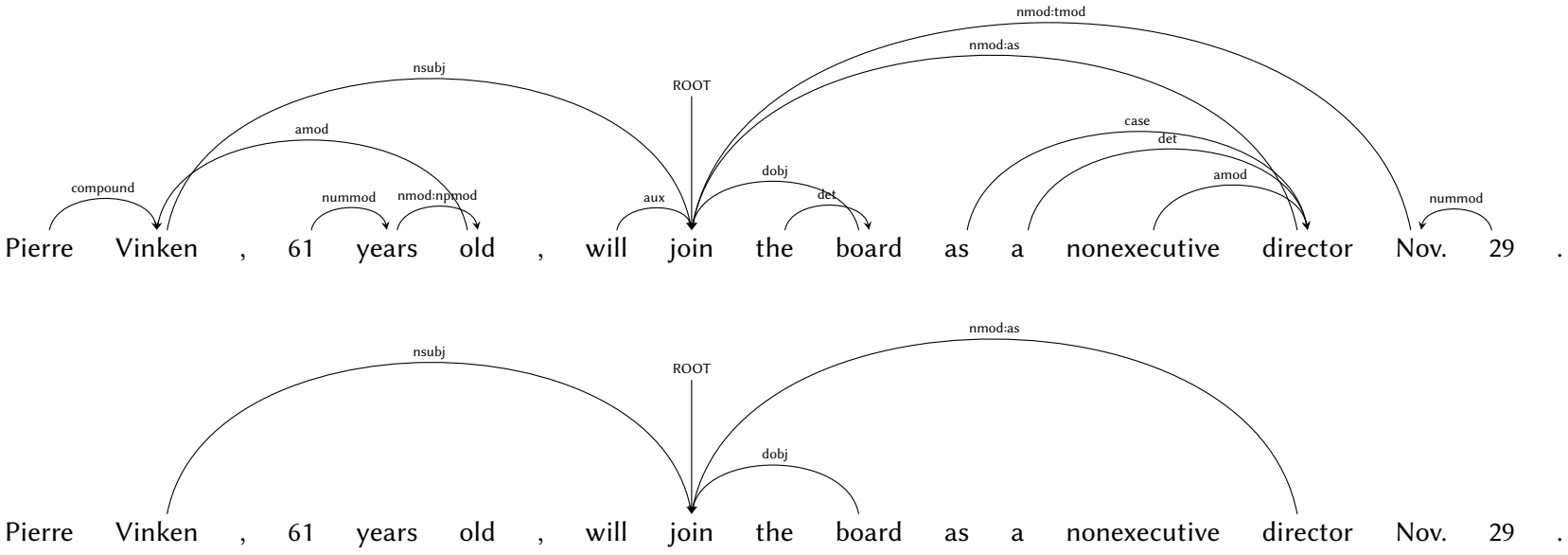
Converting Dependency Parses to Fragments (1)

- A fragment is initiated for each word tagged as a Verb (VB*).
- The identified head-verb together with the label *root* are pushed into this fragment.
- Any word with a syntactic relation to the head together with its dependency label are also pushed to the fragment.
  - Words indirectly related to the head (through VP conjunction and disjunction relations) are added, too.

Converting Dependency Parses to Fragments (2)

- Verb particles are attached to their head.
- Passive verbs are relabeled by swapping their subject (*nsubjpass* $\mapsto$ dobj) and object (*nmod:agent* $\mapsto$ subj).
- Conjunctions, functional relationships, and some other relations between content words are removed:
- Fragments are *sorted* and sealed by appending the special symbol *EOS* as the last item.

Converting Dependency Parses to Fragments (3): Example



For the dependency structure shown here, we generate the following terminal sequence:

Example

join Vinken-:-1-:-nsubj board-:-10-:-dobj director-:-14-:-nmod:as EOS

Fragments to Grammar

- Input fragments are scanned to form T_v and T_n (i.e., the terminal symbol set).
- The maximum number of latent frames and semantic roles are assumed.
- A CFG grammar is generated such that each fragment can be parsed to any latent frame, and each of its arguments to any of latent semantic roles.

Training

- Production rules are weighted randomly such that for each nonterminal T at the LHS , $\sum w(T) = 1$.
- The weights are recalculated using the inside-outside algorithm (EM) to maximize the log-likelihood of the grammar when parsing a given set of fragments.
- For faster convergence, we alternate between the E-step and the M-step in a stepwise manner (e.g., for every 1000 fragments).

Generating Output

- Once a model is trained (e.g., after 10 to 20 iterations or when changes in likelihood is negligible), we use the estimated weights to form a PCFG.
- Input fragments are parsed using the obtained PCFG (e.g., using the CYK algorithm).
- The best parse (i.e., one with the highest probability) is found and chosen to assign frame and semantic role information to the input fragment.

Example

join.f2
Vinken:Role1 board:Role7 director:Role2

Evaluation: The Dataset

- We perform our experiments on the WSJ sections of PTB.
- Namely, we create our main gold dataset using manual annotations available in the Prague Semantic Dependencies dataset (PSD). (Oepen et al., 2016; Cinkova et al., 2012).
- For verb frames, PSD provides fine grained annotations for both sense/frame and semantic roles.
- Additionally, where it is possible we map PSD annotation instances to:
  - Propbank sense and argument groupings (Palmer et al., 2005) using EngVallex (Cinková et al., 2014);
  - FrameNet (Baker et al., 1998), and VerbNet (Kipper et al., 2000) groupings derived from automatic analysis of the SemLink 2.3c dataset (Bonial et al., 2013);
  - OntoNote sense grouping of verbs derived from the SemLink 2.3c dataset.

Evaluation: The Dataset

	#Sent.	#Inst.	#LH	#Frames	#IPHS	#Arg.	#AAPS	#AT
SDP_O	35,910	89,912	3569	5,448	12.13	220,837	3.77	77
SDP_T[1]	32,195	74,023	2,409	4,223	16.48	182,365	4.35	75
Propbank	33,150	71,881	2,758	3,640	17.57	127,182*	2.11	20
VerbNet	21,502	30,766	1,254	1,473[2]	20.41	51,276	1.90	27
FrameNet	17,282	22,862	796	1,012	21.73	21,161	1.14	189
OntoNote	11,807	18,891	1,165	1,871	8.79	na	na	na

Statistics of the datasets in our experiments: #LH denotes the number of distinct lexical heads; #IPHS is the average number of instances per sense and the #APS is the average number of arguments per sense. #AT denotes the number argument classes.

[1] We remove copulas as well as gerunds (i.e., words tagged as VBG) with an adjectival function from SDP_O.

[2] Verb classes

Evaluation: Fragment Generation (1)

- Ideally, syntactic verb–argument structures derived from the enhanced universal dependency parses have a similar topology of semantic frames in our gold data.[3]
- In practice, there are mismatches between syntactic verb–argument structures and semantic ones, e.g., due to errors from automatic processing, annotation guidelines, annotation errors, etc.
- We measure these inconsistencies by comparing syntactic structures (obtained from automatic dependency parses) and our gold semantic structures.
  - We choose precision (P), recall (R), and their harmonic mean (F1-score=$\frac{2\times P\times R}{P+R}$) to measure the performance.

$$P = \frac{true\ positive}{true\ positve + false\ positive},$$

$$R = \frac{true\ positive}{true\ positve + false\ negative}.$$

[3] Given that enhanced universal dependency parses link content words.

Evaluation: Fragment Generation

	#++	#+-	#-+	P	R	F1
Head	72,721	1,302	2,282	0.969	0.982	0.976
Argument	136,515	45,850	42,950	0.761	0.748	0.754
Head+Arg[4]	136,515	43,666	39,140	0.776	0.767	0.742

- Evaluating Fragment Generation: For Head+Arg (considering head and argument together), we report P, R, and F1 micro averaged per head.
- Note that the last line set an upper bound for the performance.
- For clustering evaluation, we focus on the intersection of generated fragments and the gold frames.

[4] Note that here, arguments of unidentified heads are removed from the argument set.

Evaluation: Clustering: Figures of Merit (1)

- The aim is to define and measure similarity between automatically generated clusters and the gold data groupings: Extrinsic Clustering Evaluation Metrics.

- The Set matching measures of Purity (Pu), Inverse-Purity (iPu), and their Harmonic Mean:
  - These measures assume a mapping between system generated clusters and gold categories.
  - Pu finds the most common category (i.e., maximum precision) in each cluster and takes a weighted average:

$$Pu = \sum_i \frac{|C_i|}{N} \max_j \frac{|C_i \cap G_j|}{C_i} = \frac{1}{N} \sum_i \max_j |C_i \cap G_j|,$$

    where N is the number of instances being clustered.
  - Inversely, iPu finds mappings that maximizes recall.
  - The harmonic mean is given by $\frac{2 \times Pu \times iPu}{Pu + iPu}$.

Evaluation: Clustering: Figures of Merit (2)

- Pu penalizes the presence of noise in clusters (items from mixed categories), but it does not give credit for clustering items from the same category.
- iPu rewards grouping items together without penalizing of the presence of noise (mixed items).
- We can reach maximum Pu=1.0 by putting each instance in a cluster, and the maximum iPu=1.0 by putting all instances in one cluster.
  - Given the Zipfian nature of the distribution of instances in categories (e.g., word senses), putting all instances in one cluster results in high Pu, too!
- Also, for Pu and iPu, only clusters with the majority of items from each category matters: e.g., there is no difference between Pu and iPu of clustering of 8 items in three clusters of size 5, 2, and 1 clusters or in two clusters of size 5, and 3![5]

[5] For both, Pu = 1.0, iPu = 0.77, and f=0.769.

Evaluation: Clustering: Figures of Merit (3)

- Simply put, PU and iPu are 'cluster-centric' measures.
- To give a better picture of our method's performance, we report B-Cubed Precision, Recall, and mean.
- B-Cubed measure is item centric: Instead of focusing on a system-generated clusters (i.e., one cluster at a time), for each item t in a cluster, it measures the presence or absence of other items from the same class of t across all clusters.
- The final P_b and R_b are then given by $P_b = \sum_i^N w_i * P_i$, and $R_b = \sum_i^N w_i * R_i$, where N is the number of instances being clustered and w_i is a weight assigned to each instance. Following Bagga and Baldwin (1998), we use constant $w_i = \frac{1}{N}$.
- For instance, B-Cubed measures for 3+5 and 5+2+1 are respectively: $P_b = 1.0$, $R_b = 0.53$ and B-Cubed = 0.69 vs. $P_b = 1.0$, $R_b = 0.46$ and B-Cubed = 0.63.[6]

[6] see Amigó et al. (2009) for further analysis of these metrics.

Results: Frame/Sense Discrimination

- Results for PSD as Gold Data; $|F| = 2, |R| = 3$.

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
B-1CPH	56725	1973	745	78.43	99.66	87.78	71.08	99.5	82.92	94.72	99.98	90.09	94.78
B-All1C	56725	1973	1	10.86	100	19.6	1.48	100	2.91	0.0	100	1.48	2.91
B-Rnd-2	56725	1973	1420	78.58	55.35	64.95	71.39	51.55	59.87	89.68	50.05	90.1	64.36
B-init	56725	1973	1375	79.06	72.79	75.79	71.94	65.63	68.64	91.03	83.82	92.66	88.02
Sys-EM	56725	1973	896	79.37	95.22	86.58	72.1	93.35	81.36	94.14	99.07	91.81	95.3

Results: Frame/Sense Discrimination

- Results for VerbNet Classes as Gold Data; $|F| = 2, |R| = 3$.

Name	Inst.	Gold	Clust	*Pu*	*iPu*	*f*	P_b	R_b	B_f	V-m	P	R	f
B-1CPH	25538	572	405	93.74	100	96.77	91.56	100	95.59	98.53	100	99.05	99.52
B-All1C	25538	572	1	20.21	100	33.62	4.61	100	8.82	0.0	100	4.61	8.82
B-Rnd-2	25538	572	791	93.88	54.52	68.98	91.68	51.06	65.59	92.03	50.05	99.05	66.5
B-init	25538	572	760	94.05	75.16	83.55	91.88	68.4	78.42	93.98	87.62	99.34	93.11
Sys-EM	25537	572	498	94.48	96.1	95.28	92.28	94.52	93.39	97.77	99.59	99.25	99.42

Results: Frame/Sense Discrimination

- Results for FrameNet as Gold Data; $|F| = 2, |R| = 3$.

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
B-1CPH	19987	401	245	87.37	100	93.26	84	100	91.3	97.28	100	84.9	91.84
B-All1C	19987	401	1	3.6	100	6.95	0.94	100	1.87	0	100	0.94	1.87
B-Rnd-2	19987	401	477	87.65	54.61	67.29	84.2	51.08	63.59	91.2	50.27	85.03	63.18
B-init	19987	401	463	87.62	70.13	77.9	84.25	62.54	71.79	92.36	59.1	82.34	68.81
Sys-EM	19987	401	309	87.84	94.43	91.02	84.51	92.23	88.2	96.25	94.95	87.31	90.97

Results: Frame/Sense Discrimination

- Results for Propbank Groupings as Gold Data; $|F| = 2, |R| = 3$.

Name	Inst.	Gold	Clust	*Pu*	*iPu*	*f*	P_b	R_b	B_f	V-m	P	R	f
B-1CPH	52465	951	534	90.39	100	94.95	86.83	100	92.95	97.76	100	97.02	98.49
B-All1C	52465	951	1	13.51	100	23.8	2.19	100	4.29	0.0	100	2.19	4.29
B-Rnd-2	52465	951	1068	90.5	53.93	67.58	86.93	50.86	64.18	92.04	50.04	97.02	66.03
B-init	52465	951	1045	90.83	72.87	80.87	87.3	65.89	75.1	93.55	85.76	97.91	91.43
Sys-EM	52464	951	664	91.15	95.25	93.16	87.65	93.48	90.47	96.99	99.12	97.76	98.44

- Obviously, the system recognizes some useful patterns in the data and discriminates between word senses.
- Comparing these results with those reported in Manandhar et al. (2010) indicates the effectiveness of the method.[7]

[7] Near-Future work!

Results: The $|F|$ Hyper-Parameter

$\|F\|$	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
2	1973	896	79.37	95.22	86.58	72.1	93.35	81.36	94.14	99.07	91.81	95.3
5	1973	1034	79.92	92.22	85.63	72.81	89.07	80.12	93.72	98.16	91.94	94.95
7	1973	1707	80.66	73.67	77	73.92	66.44	69.98	91.04	91.69	94.77	93.21
10	1973	1923	81.9	70.07	75.52	75.39	61.36	67.66	90.39	78.26	94.51	85.62

- Large $|f|$ leads to lower performance; although, splitting clusters obtained from large $|F|$ may result in better performance.

Results: Argument Clustering (1)

- Result for PSD as gold data, $|F| = 2$ and $|R| = 3$.

Name	Inst.	Gold	Clust	*Pu*	*iPu*	*f*	P_b	R_b	B_f	V-m	P	R	f
B-Rnd3	101624	71	3	33.72	34	33.86	23.14	33.37	27.33	0.04	33.34	23.14	27.31
B-1CPGR	101624	71	81	73.05	72.56	72.8	63.42	59.66	61.48	54.21	68.23	80.63	73.92
B-All1C	101624	71	1	33.68	100	50.39	23.13	100	37.58	0.0	100	23.13	37.58
B-init	101624	71	3	40.65	47.6	43.85	24.74	37.84	29.92	3.87	36.57	25.1	29.77
Sys-EM	101625	71	3	55.2	74.82	63.53	35.24	62.25	45	19.95	60.83	32.7	42.53

Results: Argument Clustering (2)

- Result for Propbank as gold data, $|F| = 2$ and $|R| = 3$.

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
B-Rnd3	71455	19	3	47.62	33.75	39.5	36.05	33.35	34.65	0.02	33.34	36.05	34.64
B-1CPGR	71455	19	70	77.94	71.55	74.61	66.36	57.93	61.86	41.42	60.45	72.12	65.77
B-All1C	71455	19	1	47.62	100	64.52	36.05	100	52.99	0.0	100	36.05	52.99
B-init	71455	19	3	50.38	44.32	47.16	37.28	35.8	36.52	1.72	35.93	37.57	36.73
Sys-EM	71456	19	3	66.72	74.92	70.58	49.23	60.52	54.29	19.96	59.36	48.01	53.08

Results: Argument Clustering (3)

- Result for PSD as gold data, $|F| = 2$ and $|R| = 7$.

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
B-Rnd7	101624	71	7	33.74	14.89	20.66	23.14	14.35	17.71	0.11	14.29	23.14	17.67
B-1CPGR	101624	71	81	73.05	72.56	72.8	63.42	59.66	61.48	54.21	68.23	80.63	73.92
B-All1C	101624	71	1	33.68	100	50.39	23.13	100	37.58	0.0	100	23.13	37.58
B-init	101624	71	7	50.97	33.55	40.47	30.73	22.01	25.65	9.33	22.47	33.25	26.82
Sys-EM	101625	71	7	65.35	75.96	70.26	49.97	63.95	56.1	41.31	68.41	50.72	58.25

Results: Argument Clustering (4)

- Result for Propbank as gold data, $|F| = 2$ and $|R| = 7$.

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
B-Rnd3	71455	19	3	47.62	33.85	39.57	36.05	33.35	34.65	0.02	33.34	36.05	34.64
B-1CPGR	71455	19	70	77.94	71.55	74.61	66.36	57.93	61.86	41.42	60.45	72.12	65.77
B-All1C	71455	19	1	47.62	100	64.52	36.05	100	52.99	0.0	100	36.05	52.99
B-init	71455	19	7	59.1	32.62	42.04	43.44	21.1	28.41	7.96	20.79	45.11	28.46
Sys-EM	71455	19	7	74.7	80.69	77.58	59.33	68.02	63.37	34.53	68.95	59.08	63.63

Results: Argument Clustering (5)

- Result for Propbank as gold data, $|F| = 2$ and $|R| = 7$.

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
B-Rnd3	71455	19	3	47.62	33.85	39.57	36.05	33.35	34.65	0.02	33.34	36.05	34.64
B-1CPGR	71455	19	70	77.94	71.55	74.61	66.36	57.93	61.86	41.42	60.45	72.12	65.77
B-All1C	71455	19	1	47.62	100	64.52	36.05	100	52.99	0.0	100	36.05	52.99
B-init	71455	19	7	59.1	32.62	42.04	43.44	21.1	28.41	7.96	20.79	45.11	28.46
Sys-EM	71455	19	7	74.7	80.69	77.58	59.33	68.02	63.37	34.53	68.95	59.08	63.63

Results: Argument Clustering (6)

- Result for VerbNet as gold data, $|F| = 2$ and $|R| = 7$.

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
B-Rnd7	30083	27	7	38.59	15.28	21.9	24.57	14.35	18.11	0.11	14.3	24.57	18.08
B-1CPGR	30083	27	53	64.95	78.78	71.2	44.87	68.9	54.35	36.49	77.09	45.97	57.59
B-All1C	30083	27	1	38.59	100	55.69	24.55	100	39.42	0.0	100	24.55	39.42
B-init	30083	27	7	50.33	37.11	42.72	30.56	23.87	26.81	8.17	24.67	32.95	28.21
Sys-EM	30084	27	6	62.41	82.17	70.94	41.89	71.59	52.85	28.51	75.67	42.61	54.52

Results: Argument Clustering (7) – Detailed

- PSD as Gold Data, $|F| = 2$ and $|R| = 7$; the metric is the harmonic mean of Pu and iPu.

Role[8]	**B1CG**	$\mathcal{B}$	**OM**	$\mathcal{C}$	size
PAT-arg	0.818	13	0.896	6	34226
ACT-arg	0.971	11	0.886	6	33913
TWHEN	0.577	10	0.531	7	4413
RHEM	0.79	7	0.79	6	3157
EFF-arg	0.684	10	0.723	6	3039
LOC	0.936	7	0.649	6	2707
ADDR-arg	0.629	9	0.849	6	2652
MANN	0.85	8	0.85	7	1903
CPHR-arg	0.922	8	0.939	6	1503
DIFF	0.927	7	0.987	6	1458
AIM	0.544	7	0.891	4	1007
ORIG-arg	0.985	5	0.972	6	964
DIR3-arg	0.885	6	0.782	6	940
COMPL	0.682	8	0.754	5	898
EXT	0.772	7	0.773	6	750

Role	**B1CG**	$\mathcal{B}$	**OM**	$\mathcal{C}$	size
CAUS	0.774	7	0.782	5	522
MEANS	0.799	6	0.722	6	471
COND	0.908	9	0.931	4	462
ACMP	0.904	6	0.885	6	454
DIR3	0.873	7	0.77	6	408
REG	0.884	6	0.775	4	408
TTILL	0.717	5	0.631	6	374
THO	0.916	5	0.916	6	364
THL	0.701	5	0.698	5	362
DPHR-arg	0.79	7	0.941	5	319
PREC	0.888	3	0.888	4	317
ATT	0.919	7	0.921	5	291
LOC-arg	0.93	6	0.632	6	288
BEN	0.924	6	0.864	4	274
EXT-arg	0.88	7	0.978	5	252

[8] See `https://ufal.mff.cuni.cz/pcedt2.0/en/functors.html` for tag description.

Results: Argument Clustering: The $|R|$ Hyper-Parameter

- Result for PSD as gold data, $|F| = 2$ and $|R| = 11$.
- Increasing $|R|$ beyond 7 has a diverse effect.

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
B-Rnd11	101624	71	11	33.87	10.62	16.17	23.14	10.06	14.02	0.13	10	23.14	13.97
B-1CPGR	101624	71	81	73.05	72.56	72.8	63.42	59.66	61.48	54.21	68.23	80.63	73.92
B-All1C	101624	71	1	33.68	100	50.39	23.13	100	37.58	0.0	100	23.13	37.58
B-init	101624	71	11	44.91	20.87	28.49	27.84	13.15	17.86	7.24	12.3	28.57	17.2
Sys-EM	101625	71	11	68.16	42.97	52.71	57.08	33.71	42.38	42.32	30.06	65.16	41.14

Some results from currently running experiments

- Replace unification-based terminal to non-terminal mappings with distributional similarities:
  - Given terminal t in input terminal sequence, instead of only counting $T \mapsto_\theta t$ with weight θ, allow unification of all $T \mapsto_\eta x$ with the weight $\eta e^{correlation(t,x)}$.

- Results Using PSD as gold data; $|F| = 2$ and $|R| = 3$:
  - Sense Discrimination

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
EM (Unif.)	56725	1973	896	79.37	95.22	86.58	72.1	93.35	81.36	94.14	99.07	91.81	95.3
EM (PoP-T)[9]	56725	1973	916	78.57	98.37	87.36	71.33	97.35	82.33	94.42	99.09	90.29	94.48

  - Role Induction

Name	Inst.	Gold	Clust	Pu	iPu	f	P_b	R_b	B_f	V-m	P	R	f
EM (Unif.)	101625	71	3	55.2	74.82	63.53	35.24	62.25	45	19.95	60.83	32.7	42.53
EM (PoP-T)	101625	71	3	33.68	98.84	50.24	23.2	97.74	37.5	0.53	98.52	23.33	37.72

[9] Slow but steady convergence; correlation is computed using r^2 coefficient of determination.

Conclusion & future work

- Obviously the model recognizes 'some' patterns similar to frames (performances better than several baselines, except one).
- An intuitive interpretation of these learnt patterns remains an open question:
  - Problems due to the use of extrinsic clustering evaluation metrics.
  - Alternative definitions for performance and thus evaluation, suggestions?!
- And, after all, the gold data is not that gold:
  - "The Senate ... still is expected to join the House in voting to kill the law, which ...":
  - GOLD: #22048013 17 join.ev-w1777f1 Senate-:-1-:-ACT-arg house-:-19-:-PAT-arg vote-:-21-:-TWHEN
  - The automatic clustering, however, results in 'vote' to be clustered with items which are oftentimes annotated as *EFF-arg* in the gold set.

Conclusion & future work

- Altering the underlying assumptions about the proposed mixture multinomial probabilistic model:
  - Altering the CFG (i.e., random variables and their structural dependencies), additional feature set, etc.
  - Or, altering the model type (e.g., from mixture multinomial to mixture Gaussian), and, consequently parameter types, and their estimation method.
- Altering the learning (parameter estimation) procedure:
  - The well known Split–Merge method.
  - Learning one thing at a time, and then learning more about what has been learnt.
  - Recursive application of the method, e.g. to replace simple word terminal symbols with frames learnt in a previous procedure.

References I

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486.

Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 79–85, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bonial, C., Stowe, K., and Palmer, M. (2013). Renewing and revising semlink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9 – 17, Pisa, Italy. Association for Computational Linguistics.

Cheung, J., Poon, H., and Vanderwende, L. (2013). Probabilistic Frame Induction. In *Proceedings of NAACL-HLT*, pages 837–846.

Cinková, S., Fučíková, E., Šindlerová, J., and Hajič, J. (2014). EngVallex - english valency lexicon. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

References II

Cinkova, S., Mikulova, M., Mladova, L., Nedoluzko, A., Pajas, P., Panevova, J., Semecky, J., Sindlerova, J., Uresova, Z., Zabokrtsky, Z., Semecky, J., Sindlerova, J., Toman, J., Uresova, Z., and Zabokrtsky, Z. (2012). Annotation of english on the tectogrammatical level: Reference book.

Kipper, K., Dang, H. T., and Palmer, M. (2000). Class-based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press.

Manandhar, S., Klapaftis, I. P., Dligach, D., and Pradhan, S. S. (2010). Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 63–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

Oepen, S., Kuhlmann, M., Miyao, Y., Zeman, D., Cinkova, S., Flickinger, D., Hajic, J., Ivanova, A., and Uresova, Z. (2016). Towards comparability of linguistic graph banks for semantic parsing. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.

References III

Schuster, S. and Manning, C. D. (2016). Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).