Enriching Texts with Interpretable Topics of Propositions

Henry Anaya-Sánchez and Anselmo Peñas
{henry.anaya, anselmo}@lsi.uned.es

Advances in Distributional Semantics Workshop (Co-located with IWCS 2015)

NLP & IR Group, UNED
Juan del Rosal 16,
28040 Madrid, Spain
http://nlp.uned.es

Motivation

- We consider the task of **semantically enriching** texts at the sentence level using **proposition-based topics** that model the **main events** underlying the texts.
- The aim is to obtain semantically deeper representations of texts to be applied for further text analysis; e.g., to predict basic actions, events or sentences in a text document.

Our approach

- A general methodology that extends the framework of PTM to allow mapping natural language sentences to a topic-like representation of events based on distributions of propositions; where each distribution is deemed to be a human interpretable abstraction useful to describe the main actions involved in the sentences.
- Provides an enriched representation for sentences that describes their main contents, even though the propositions in the descriptions do not explicitly appear in the texts.

Basic components

- A PS (i.e., a collection of propositions gathered from a reference collection of unlabeled texts) $S = \{s_i\}_i$ representing the BKB from which to learn the statistical models to base the enrichment of an input text.

- A set of classes $C = \{c_1, \dots, c_{|C|}\}$ such that:
  instances of predicate argument $A_{r,i}$ can be represented by $C_{r,i} \subseteq C$.

- A class-based selectional preference model Γ that maps each predicate r to a discrete distribution of class tuples that model the stochastic generation of individual propositions with the form $s = r(a_1, \dots, a_{arity(r)})$:

$$\Gamma(r) \vdash (C_{r,1} \times \cdots \times C_{r,arity(r)} \to (0,1])$$

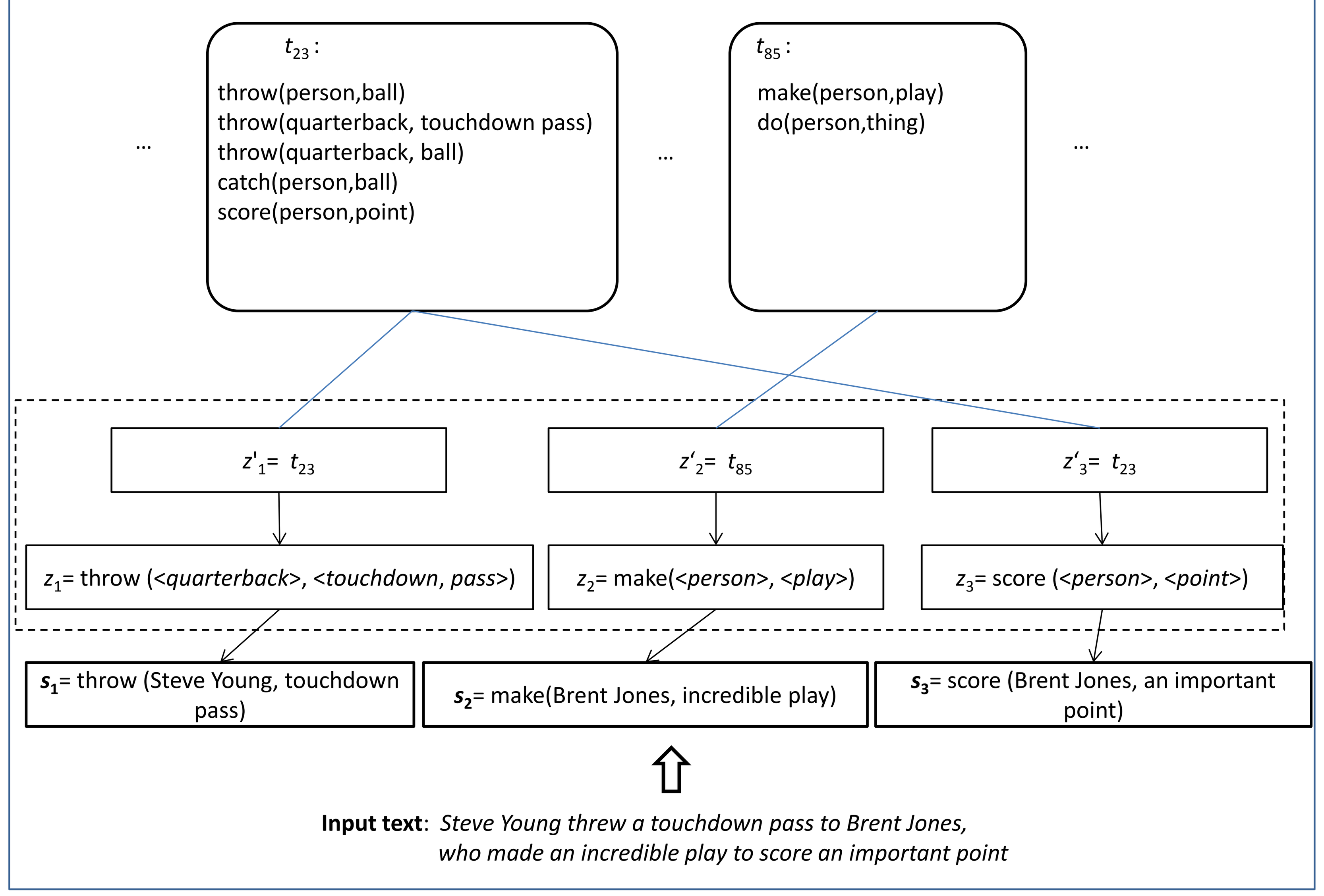
To support interpretability, classes are represented by means of common nouns (specifically, nominal phrases)

$$p\left(s = r(a_1, \dots, a_{arity(r)})\right) = \sum_{k=1}^{|\Gamma(r)|} p(z = \Gamma(r)_k) \prod_{i=1}^{arity(r)} p(a_i | z(i))$$

$$\Gamma(throw) = \begin{cases} (quarterback, pass) & 0.54 \\ (quarterback, interception) & 0.21 \\ (quarterback, ball) & 0.07 \\ (person, ball) & 0.06 \\ (group, ball) & 0.05 \\ \vdots & \vdots \\ (person, interception) & 0.002 \end{cases}$$

$$\mathrm{p}(s = throw(\boldsymbol{Young}, \boldsymbol{ball})) = 0.54\, \mathrm{p}(\boldsymbol{Young}|quarterback)\, \mathrm{p}(\boldsymbol{ball}|pass) + \vdots + 0.002\, \mathrm{p}(\boldsymbol{Young}|person)\, \mathrm{p}(\boldsymbol{ball}|interception)$$

Methodology



Enriching argument instances with classes

$$s_i = r_i(a_{i,1}, \dots, a_{i,arity(r_i)})$$

Label $z_i = r_i(z_i[1], \dots, z_i[arity(r_i)])$

where $z_i = \mathrm{argmax}_z\ \mathrm{p}(z|s_i)$

$$\mathrm{p}(z = \Gamma(r_i)_k \mid s_i) \propto \mathrm{p}(z = \Gamma(r_i)_k) \prod_{j=1}^{arity(r_i)} \mathrm{p}(a_{i,j}|z[j])$$

Enriching sentences with topic-like events

- A similar generative story to that of LDA to label each z_i with topic:
$$z'_i \in T = \{t_1, \dots, t_{|T|}\}$$
but considering the collection $\{z_i\}_i$ as a single document.

- Each topic is learned as an explanation of a given class-based proposition :
$$\exists\ z_t \text{ behind each topic } t$$

- Topics are not so latent and we use a fixed $p(z|t) = p(z|z_t)$ (a stochastic mapping between class-based propositions estimated from co-occurrences).

- Constrained sampling for $z = r(c_1, \dots, c_{arity(r)})$:
  possible topics are those behind $z^* = r^*(c^*_1, \dots, c^*_{arity(r^*)})$
  where $\mathrm{PMI}(r, r^*)$ and $\mathrm{PMI}(z, z^*)$ are greater than a threshold.

Component Learning

- Classes:
$$\mathrm{p}(c|A_{r,i}) \propto \sum_{c' \in C} \sum_{a \in A^*} \mathrm{p}^*(c|c')\mathrm{p}(c'|a)\mathrm{p}(a|A_{r,i})$$
We define $C_{r,i} = \{c \mid \mathrm{p}(c|A_{r,i}) > \theta_0,\ \mathrm{PMI}(c, A_{r,i}) > \gamma_0\}$

- Class-based SPF model:
For each r, we need to infer some priors for the tuples in $C_{r,1} \times \cdots \times C_{r,arity(r)}$.
We consider a Gibbs sampling procedure that randomly assigns a class-based proposition to each $s = r(a_1, \dots, a_{arity(r)})$ according to:
$$\mathrm{p}(z = \Gamma(r)_k \mid s) \propto \frac{n_k + \alpha}{N + \alpha K} \prod_{i=1}^{arity(r)} p(a_i|z[i])$$

Experiments

We consider a collection of 30,826 New York Times articles about US football that was partitioned in 80%-20% training-test documents. The top 1500 most frequent predicates were considered.

Event interpretability
Averaged values of the Umass coherence measure obtained for the learned distributions of class-based propositions (ε=1.0e-50).

Method	n=5	n=10	n=15	n=20
HDP-baseline	-748.67	-3622.4	-9074.4	-17489.1
Our proposal	-630.39	-3020.6	-5652.6	-7368.1

Predicting propositions
Averaged values of log-likelihood obtained in the generation of test documents represented as bags of propositions.

Document size	HDP-baseline	Our proposal
[1;50]	-90.68	-22.64
[51;100]	-220.52	-84.77
[101;150]	-290.19	-102.67
[151;200]	-367.15	-113.04
[201;250]	-422.37	-89.51
any	-216.01	-75.05

References

•David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res., 3:993–1022, March.
•D. Mimno, H.M. Wallach, E. Talley, M. Leenders, and A. McCallum. 2011. Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 262–272.
•Anselmo Peñas and Eduard Hovy. 2010. Filling knowledge gaps in text for Machine Reading. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 979–987. Association for Computational Linguistics.
•Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. 2006. Hierarchical Dirichlet Processes. Journal of the American Statistical Association, 101(476):1566–1581.

Acknowledgments This work has been partially supported by MINECO (PCIN-2013-002-C02-01) and EPSRC (EP/K017845/1) in the framework of CHIST-ERA READERS project, and by project Voxpopuli (TIN2013-47090-C3-1).