Technical report for Events' labels matching computation

Karn Yongsiriwit
Telecom SudParis, UMR 5157 CNRS Samovar, France
Email: karn.yongsiriwit@telecom-sudparis.eu

I. Events' labels matching

Function $sim.\lambda_p$ computes the similarity of two labels. We use Stanford Part-of-Speech (POS) [1], [2] for stemming string and removing function words. We modify the bag-of-words similarity with label pruning technique [3] to prunes words from the longer label then measure the similarity of two labels based on pruned words.

The similar label matching between label l_1 and l_2 is defined as follows:

$$sim.\lambda_p(l_1, l_2) = \frac{\sum_{i=1}^{|\omega^1|} max_{j=1}^{|\omega^2|}(sim.w(pr_1^i, pr_2^j)) + \sum_{j=1}^{|\omega^2|} max_{i=1}^{|\omega^2|}(sim.w(pr_1^i, pr_2^j))}{2 \times min(|\omega^1|, |\omega^2|)} \quad (1)$$

Let l_1 and l_2 be the labels of the events e_1 and e_2, and $\omega^1 = tok(\lambda_1(a_1))$, $\omega^2 = tok(\lambda_2(a_2))$ are tokenized lists of words contained in the labels by POS technique. Further, $pr_1 = pru(\omega^1, \omega^2)$ and $pr_2 = pru(\omega^2, \omega^1)$ are the pruned list of words. Function $sim.w$ computes the similarity between two words using existing approaches such as Lin metric ontology matching technique [4]. We define a threshold β_l for $sim.\lambda_p$. Two labels are considered to be functionally similar iff $sim.\lambda_p(l_1, l_2) \geq \beta_l$. Let $pru : P(W) \times P(W) \rightarrow P(W)$ be a generic function. It returns a set of words extracted from its input. $pru(\lambda_p(e1), \lambda_p(e2))$ is ω^1 iff $|\omega^1| \leq |\omega^2|$, or a subset of ω^1 of size $|\omega^2|$ otherwise. The similarity scores of all word pairs, as well as the maximum score for each word are calculated in $|\omega^1|$. *pru* returns the $|\omega^2|$-top-scoring words from ω^1.

In order to illustrate our approach, we present the similarities between labels of the events in the $1^{st} - zone$ of the neighborhood context of C and $C2$ using ($sim.\lambda_p$) (see our motivating example). We define the threshold as $\beta_l = 0.5$. The results after descending sort are shown Table II. The details about the example computation are shown in Table I. Note that the word pairs with the similarity value underlined indicates the maximum similarity value among other word pairs.

We create the matching between the most similar events by the ranking the similarity values. The result are event B2 is matched to event B with $sim.\lambda_p(B2, B) = 1.000$ and event D2 is matching to event D with $sim.\lambda_p(D2, D) = 0.543$.

$$N_c^1(C, C2) = \{(B2, B, 1.000), (D2, D, 0.543)\}$$

Table I. Similarity computation of all possible neighbors pairs between events C and$C2$ in the $1^{st} - zone$ neighborhood context

		A	
	sim.w	send	e-mail
B2	process	0.000	0.000
	check	0.000	0.000
	credit	0.000	0.000
	request	0.576	0.000
	response	0.000	0.000

(a) $sim.\lambda_p(l_{B2}, l_A) = 0.288$

		B	
	sim.w	check	credit
B2	process	0.099	0.096
	check	1.000	0.085
	credit	0.335	1.000
	request	0.369	0.469
	response	0.000	0.000

(b) $sim.\lambda_p(l_{B2}, A_{label}(B)) = 1.000$

		D	
	sim.w	check	system
B2	process	0.099	0.000
	check	1.000	0.000
	credit	0.335	0.000
	request	0.369	0.000
	response	0.000	0.073

(c) $sim.\lambda_p(l_{B2}, l_D) = 0.537$

		E
	sim.w	accept
B2	process	0.000
	check	0.000
	credit	0.455
	request	0.000
	response	0.000

(d) $sim.\lambda_p(l_{B2}, l_E) = 0.455$

		F
	sim.w	reject
B2	process	0.000
	check	0.000
	credit	0.439
	request	0.000
	response	0.000

(e) $sim.\lambda_p(l_{B2}, l_F) = 0.439$

		A	
	sim.w	send	e-mail
D2	check	0.000	0.000
	paper	0.000	0.000
	archive	0.000	0.000

(f) $sim.\lambda_p(l_{D2}, l_A) = 0.000$

		B	
	sim.w	check	credit
D2	check	1.000	0.335
	paper	0.000	0.000
	archive	0.000	0.000

(g) $sim.\lambda_p(l_{D2}, l_B) = 0.668$

		D	
	sim.w	check	system
D2	check	1.000	0.000
	paper	0.000	0.085
	archive	0.000	0.000

(h) $sim.\lambda_p(l_{D2}, l_D) = 0.543$

		E
	sim.w	accept
D2	check	0.000
	paper	0.000
	archive	0.000

(i) $sim.\lambda_p(l_{D2}, l_E) = 0.000$

		F
	sim.w	reject
D2	check	0.000
	paper	0.000
	archive	0.000

(j) $sim.\lambda_p(l_{D2}, l_F) = 0.000$

Table II. Possible pairs ranked by $sim.\lambda_p$.

No.	e_1	e_2	$sim.\lambda_p(e_1, e_2)$
1	B2	B	1.000
2	D2	B	0.668
3	D2	D	0.543
4	D2	B	0.537

Using $sim.\lambda_p$, the log-based neighborhood context matching is described in Equation 2. We weight the similar pair

of events based on their labels similarity.

$$M^{k}_{sim.\lambda_p}(a,b) = \frac{\sum_{i=1}^{z} sim.\lambda_p(l_{r_{a_i}}, l_{r_{b_i}}) \times a_i \times sim.\lambda_p(l_{t_{a_i}}, l_{t_{b_i}}) \times b_i}{|\overrightarrow{e_c(a)}| \times |\overrightarrow{e_c(b)}|} \tag{2}$$

Using the label similarities from Table II, we have $N^1_c(C, C2) = \{(B2, B, 1.000), (D2, D, 0.543)\}$. So, $\overrightarrow{e_c(C)} = (w(B,C), w(C,D)) = (54, 46)$, $\overrightarrow{e_c(C2)} = ((B2, C2), (C2, D2)) = (70, 70)$ and their matching in the $1^{st} - zone$ is:

$$M^{1}_{sim.\lambda_p}(C, C2) = \frac{1.000 \times 54 \times 1.000 \times 70 + 1.000 \times 46 \times 0.543 \times 70}{\sqrt{4^2 + 6^2 + 54^2 + 46^2 + 42^2 + 13^2 + 35^2} \times \sqrt{70^2 + 70^2}} = 0.615$$

References

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)*, 2nd ed. Prentice Hall, 2008.

[2] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part–of–speech tagger," in *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 2000, pp. 63–70. [Online]. Available: http://nlp.stanford.edu/~manning/papers/emnlp2000.pdf

[3] C. Klinkmüller, I. Weber, J. Mendling, H. Leopold, and A. Ludwig, "Increasing recall of process model matching by improved activity label matching," in *BPM*, 2013, pp. 211–218.

[4] D. Lin, "An information-theoretic definition of similarity," in *ICML*, 1998, pp. 296–304.