SentiCorr: Four Step Approach for Multilingual Sentiment Analysis of Personal Correspondence

The problem

Erik Tromp and Mykola Pechenizkiy, Department of Computer Science, Eindhoven University of Technology

Language identification

1. Construct Graph (learn G from the training labelled data)

\[ G = (V, E, L, W_c, W_e) \]

\[ W_c(v, l) = \begin{cases} W_c(v, l) + 1 & \text{if } L(v) \text{ is defined} \\ 1 & \text{otherwise} \end{cases} \]

\[ W_e(l, c) = \begin{cases} W_e(c, l) + 1 & \text{if } W_e(c, l) \text{ is defined} \\ 1 & \text{otherwise} \end{cases} \]

\[ N_3(t1) = \{ \text{is, s, d, di, dit, i, te, e, en, n, t, te, est} \} \]

\[ N_3(t2) = \{ \text{is, s, t, th, thi, his, is, s, a, a, a, t, te, est} \} \]

2. Create Path (from unseen text, for which we need to do LI)

3. Match Graph and Path to determine similarity scores

\[ \forall l \in \text{Lang} : P_M(l) = \begin{cases} P_M(l) + \frac{W_e(l, c)}{\sum_e} & \text{if } \exists c \in V : L(c) = L'(c') \\ P_M(l) & \text{otherwise} \end{cases} \]

\[ \forall l \in \text{Lang} : P_M(l) = \begin{cases} P_M(l) + \frac{W_c(v, l)}{\sum_e} & \text{if } \exists e \in E : (v, w) \in V : L(v) = L'(v') \land L(w) = L'(v' + 1) \land e = (v, w) \\ P_M(l) & \text{otherwise} \end{cases} \]

LIGA: Graph-Based N-gram Language Identification on Short Texts

RBEM

Future work

Rule-Based Emission Model (RBEM)

Positive: Positive with no context good, well done
Negative: Negative with no context bad, terrible
Amplifier: Strengthen emission very much, a lot
Attenuator: Weaken emission a little, a tiny bit
Right Flip: Flips polarity to the right not, no
Left Flip: Flips polarity to the left but, however
Continuator: Progresses emission of polarity and, also
Stop: Interrupts emission punctuation

Performances: verified on Twitter, Facebook, and Hyves data; compared against traditional marketing survey data. Quantified error propagation from LI to further steps.

4-step language-aware approach;
LIGA: a novel approach for LI based on graph formalism that considers ordering of N-grams (easy to extend to incorporate other elements of grammar);
RBEM: novel approach, allowing to integrate more and more knowledge for polarity detection with patterns on terms and POS.

Work out automated relevance feedback mechanism;
Experiment with e-mail content;
Possibilities to extend to emotions;
Distinguish negative content from negative sentiment;
Learn from multiple sources, e.g. a stream of emotions extracted from facial expressions.