Noise Robustness in Aspect Extraction Task

Valentin Malykh, VK Research
Taras Khakhulin, SkolTech
Noise

- Typos, orthographical mistakes we call **noise**.

- To measure noise we take edit distance from original word to noised one.

- By original word we mean orthographically and syntactically correct word in context.

- By noised word - any word which differs from original one.
Noise Modeling

- In the real texts noise level is about 10%.

- We model noise analogously to spelling correction literature.

- There are no open corpora for languages in question with marked up spelling corrections.
Noise Modeling

\[ p \in [0, 0.3] \]

\[ B(p) \] - binomial distribution,
\[ U(1,|A|) \] - uniform distribution,
\[ |A| \] - alphabet length

Noise types:

- Current symbol deletion with probability \( B(1, p) \)
- Random symbol addition \( U(1, |A|) \) after the current one with probability \( B(1, p) \)
- Replace current symbol with random one \( U(1, |A|) \) with probability \( B(1, p) \)
- Swap two adjacent letters with probability \( B(1, p) \)
Aspect Extraction

- Aspect Mining
- Aspects could be extracted as topics

- the call quality of this phone is amazing
Aspect Extraction

t_i - distribution parameters for topic i

$p(t|d)$

<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old man</td>
<td>Олень, старуха, старуха, море, рыба, пряжа</td>
</tr>
<tr>
<td>Sea</td>
<td>Море, нёв, травяной нёв, золотая рыба</td>
</tr>
<tr>
<td>Earth</td>
<td>Землянка, пряжа</td>
</tr>
</tbody>
</table>

Документ (d): Сказка о рыбаке и рыбке

Жил старик со своей старухой у самого синего моря.
Они жили в ветхой землянке.
Ровно тридцать лет и три года.
Старик ловил неводом рыбу.
Старуха пряла свою пряжу.
Раз он в море закинул невод.
Пришел невод с одной тиной.
Он в другой раз закинул невод.
Пришел невод с травой морской.
В третий раз закинул он невод.
Пришел невод с одной рыбкой.
С непростою рыбкой, — золотою.
Как взмолится золотая рыбка? 
Голосом молвит человечным: 
«Отпусти ты, старче, меня в море.»
Attention-Based Aspect Extraction

- The model is aiming to obtain vector representations of aspects for a corpus.
- Each aspect is represented as some vector which is close to specific words (vector representations).
- Model trains matrix of vector representations for aspects.
- Model is designed to produce text vector representation based on word vector representations and so-called reconstruction which is linear combination of aspect vector representations.
- Loss function for the model is difference between two mentioned vectors.
ABAЕ model

$s$ - a sentence, $z_s$ - sentence vector representation

$a_i$ - attention weights, $y_s$ - intermediate vector representation for a sentence

e$_w$ - vector embedding for word $w$

$A$ - attention matrix

$T$ - aspect matrix

$p_s$ - weights for summing aspects

$r_s$ - reconstructed with $T$ matrix vector representation

\[
a_i = \text{softmax}(e_{w_i}^T \cdot A \cdot y_s)
\]

\[
y_s = \sum_{i=1}^{n} e_{w_i}
\]

\[
z_s = \sum_{i=1}^{n} a_i e_{w_i}
\]

\[
p_s = \text{softmax}(W \cdot z_s + b)
\]

\[
r_s = T^T \cdot p_s
\]
ABAE model

Recovered sentence embedding

Aspects matrix

Sentence embedding

Attention weights

$w_1$

$w_2$

$w_n$
Proposed Extensions

- Char embeddings which enrich existing word embeddings
- fastText word embeddings model
- RoVe word embeddings model
Word Vector Representations

- Words cannot be read by a computer like humans do, we need some numbers

- Simple representations basing on vocabulary are not enough.

\[
\begin{align*}
\text{motel} & \quad [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0] \\
\text{hotel} & \quad [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0] \\
\end{align*}
\]
Word Vector Representations

- **word2vec** - statistical co-occurrence model

- **fastText** - extension of word2vec with character n-grams

Figures belongs to T.Mikolov
Robust to Noise Vector Reps

- $||$ - concatenation
- $c_1..c_k$ one-hot vectors for symbols of word
- $n_b$ - prefix length, $n_e$ suffix length
- $\text{enc}$ - a function, left and right contexts $C_{\text{left}}$ & $C_{\text{right}}$

$$B(w) = c_1 || .. || c_{n_b}$$
$$E(w) = c_{k-n_e} || .. || c_k$$
$$M(w) = \sum_{i=1}^{k} c_i$$
$$BME(w) = B(w) || M(w) || E(w)$$

$$RoVe(w) = \text{enc}(BME(w); C_{\text{left}}, C_{\text{right}})$$
Robust to Noise Vector Reps
Robust to Noise Vector Reps

- Vector Rep for word «abbreviation»
- Left and right contexts are prey states of enc
Model Training

\[ L(x) = \log\left( \sum_{i \in C} e^{-s(x,w_i)} \right) + \log\left( \sum_{j \not\in C} e^{s(x,w_j)} \right) \]

- Negative sampling loss
Corpus

Citysearch contains 50000 review of New-York restaurants

These reviews has been marked up with such categories:

- Food
- Price
- Service
- Ambience
- Anecdotes
- Miscellaneous
We take a subset of Citysearch corpus, with reviews containing only one category from the list.

- The model extracts aspects from the corpus, these aspects then marked up by categories.
- The model extracts aspects from a text and top-aspect is taken into account.
- A category of this aspect then compared to existing one by the means of F1
Results

- The metric for the experiments is F1
- RoVe model shows the best robustness
Conclusion

- The original ABAE model is not robust to noise

- We presented several model extensions which are robust to noise

- For the future work we see the direction of testing on other languages and other state of the art models for aspect extraction
References

- Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems 2013 (pp. 3111–3119).
- Ю. В. Рубцова. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015, №1(109), –С.72–78
Thank You for Your Attention!
Word2Vec

\[ L = \frac{1}{N} \sum_{i}^{N} \ln(p(w_i|C(w_i))) \rightarrow \text{max} \]

\[ p(w_i|C(w_i)) = \text{softmax} \left( \sum_{w_k \in C(w_i)} v_{w_k}^T u_{w_i} \right) \]

\( w_i \) - in context C; \( v, u \) - word vectors
Metric $F_1$

Metric $F_1$ - is a harmonic mean of precision and recall of a classifier

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$
Original ABAE model results