Automated Snippet Generation for Online Advertising

S. Thomaidou[1], I. Lourentzou[1, 2], P. Katsivelis-Perakis[1,3], M. Vazirgiannis[1,4]
[1]Athens University of Economics and Business, [2]University of Illinois at Urbana-Champaign, [3]Harokopion University of Athens, [4]Ecole Polytechnique, Paris

Motivation

- Advertise products, services or brands alongside the search results
- Recently smaller displays on devices like tablets and smartphones have imposed the need for **smaller ad texts**
- Produce a **small comprehensive ad** while maintaining at the same time relevance, clarity, and attractiveness
- Provide an efficient solution for large websites or e-shops

Proposed Methodology

1. **Information Extraction** for mining the most important product or service keywords
2. **Sentiment Analysis** for keeping the most positive phrases that will have a good impact on the product image
3. **Natural Language Generation** for constructing a good form of the final ad-text sentences

Information Extraction

- Modified Pointwise Mutual Information Function based on Ganesan et al. 2012

$$pmi(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

Original PMI function measures strength of association between words

$$pmi'(w_i, w_j) = \log \frac{p(w_i, w_j)c(w_i, w_j)}{p(w_i)p(w_j)}$$

Rewards well associated words with high co-occurrence

- How well a phrase represents the given webpage - For each n-gram is the sum of the modified pointwise mutual information of every bigram that this n-gram contains, normalized by the total unigram length

$$Representativeness(w_i, \ldots, w_n) = \frac{1}{n} \sum_{i=0, j<i}^{n} pmi'(w_i, w_j)$$

Sentiment Analysis

- Usage of **Amazon Sentiment Dataset** - Determine the contextual polarity of a phrase
- Select the top *k* informative words as features, by measuring the Information Gain of each word
- Using all words as features provided better accuracy in the test set than choosing the top *k* informative words
- Remove all extracted keyphrases that were classified by our model as negative

Content Determination

- Choose only the best n-grams - N-grams that contain only verbs (in any form or tense) are removed from the n-gram list
- Use of Stanford Part-of-speech Tagger to eliminate n-grams of low readability - n-grams that contain sequences of five or more nouns in a row might have been erroneously extracted from the landing page

Sentence Template

- Build permutations of the best n-grams in order to add them to the final representation, assuring the limitation of 70 chars
- Two possible templates – necessary slots and value of price if it is provided
  a. <feature set> <price>
  b. <product name> with <feature set> <price>

Information Ranking Function

- Adoption of the formula for the two previous templates:
  a. $I_i(c) = \frac{1}{n+\frac{70}{l(c)}} \sum_{i=0}^{n} s_i$
  b. $I_i(c) = \frac{1}{n+\frac{70}{l(c)}} \max_{0<i<n} s_i + \sum_{i=0}^{n} s_i$
- Measures Information Gain - Penalizes little utilization of space

Readability Ranking with an Advertising Language Model

- **Advertising Language Model** of 47.984 unique ads obtained from major search engines (queries from Google Products Taxonomy)
- Feed SRILM with the above snippets – LM based on trigrams
- For each candidate keep its logarithmic probability as the value of the Readability Ranking Function, which is an indication of the likelihood that a given candidate will occur

Generated Promotional Text

Method	Product Name	Snippet
IE+NLG	VIZIO ESeries HDTV	***VIZIO ESeries with effective refresh rate, Low Price Guarantee***
IE+NLG+SA	Fuji Film Finepix JX580	***Artistically enliven photos, instantaneously increases shutter speed***

Evaluation of Advertising Text Criteria

Method	Attractiveness	Clarity	Relevance	Harmonic Mean
IE	0.387	0.667	0.660	0.538
IE+HG	0.253	0.693	0.517	0.410
IE+SA	0.433	0.697	0.680	0.575
IE+HG+SA	0.293	0.647	0.550	0.443
IE+NLG	**0.527**	**0.854**	**0.943**	**0.726**
IE+NLG+SA	**0.593**	**0.850**	**0.937**	**0.763**
IE+CP	0.257	0.617	0.423	0.381

Further Challenges

- Enrichment of the corpus for a more complete Language Model
- Classification of products and services
- Evaluation using Click-through Rate (CTR) from advertising campaigns

CIKM 2013

The research of S. Thomaidou is co-financed by the European Union (ESF) and Greek national funds via Program Education and Lifelong Learning of the NSRF - Program: Heracleitus II.

Prof. M. Vazirgiannis is partially supported by the DIGITEO Chair grant LEVETONE in France.

European Union European Social Fund
OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING investing in knowledge society
MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS, CULTURE & SPORTS MANAGING AUTHORITY
NSRF 2007-2013 programme for development EUROPEAN SOCIAL FUND
Co-financed by Greece and the European Union