Supervised Matching of Comments with News Article Segments

Dyut Kumar Sil
Dept CSA, IISc, Bangalore
dyutkumar@csa.iisc.ernet.in

Srinivasan H Sengamedu
Yahoo! Labs, Bangalore
shs@yahoo-inc.com

Chiranjib Bhattacharyya
Dept CSA, IISc, Bangalore
chiru@csa.iisc.ernet.in

ABSTRACT

Comments constitute an important part of Web 2.0. In this paper, we consider comments on news articles. To simplify the task of relating the comment content to the article content the comments are about, we propose the idea of showing comments alongside article segments and explore automatic mapping of comments to article segments. This task is challenging because of the vocabulary mismatch between the articles and the comments. We present supervised and unsupervised techniques for aligning comments to segments the of article the comments are about. More specifically, we provide a novel formulation of supervised alignment problem using the framework of structured classification. Our experimental results show that structured classification model performs better than unsupervised matching and binary classification model.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Information Storage and Retrieval

General Terms

Algorithms

Keywords

Comments, Structured Classification, Enrichment, ESA

1. INTRODUCTION

User generated content is the central theme of Web 2.0. While there are several forms of user generated content like blogs, photo/video uploads, reviews, etc., comments are the primary form of user interaction in several sites. Comments form a light-weight mechanism for user participation as they are primarily reactive. Recently [3] proposed mapping comments to article segments to simplify readers' task of relating comment content to article content.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

In this paper we consider a variation of the problem proposed in [3]. More specifically, while [3] addressed the unsupervised matching problem, we consider supervised matching. We state the problem formally as follows.

Problem Statement: Let an article A be characterized by the set of segments $S(A) = \{s_1, \ldots, s_{n_s}\}$, where a segment s_k could be a paragraph, sentence, etc. and they form a partition of the article A. The comments associated with article A are denoted $C(A) = \{c_1, \cdots, c_{n_c}\}$. Given a set of matching (comment, segment) pairs, the goal is to design a learning machine which, when presented with a article A and its associated comments, $C(A)$, correctly identifies for each comment, $c \in C(A)$, the related segment or segments in $S(A)$. □

Technical Contributions.

Our technical contributions directly address the above challenges and are two fold. In a machine learning scenario, the key elements are the features as well as the classification technique. We address and contribute on both fronts.

Supervised Classification: We provide two different formulations for the supervised comment alignment problem – the well-known binary classification as well as the less familiar structured classification. We show that structured classification outperforms binary classification.

Enriched Topic Features: Extending the ideas of enrichment presented in [3], we propose the use of Explicit Semantic Analysis [1] for the comment alignment task.

The problem of retrieving article segments for comments has the distinct flavor of traditional information retrieval where segments can be considered documents and the comments can be considered queries. We show through our experiments that the proposed techniques outperform two traditional IR representatives, viz., Lucene and Indri. We also show that the notion of enrichment complements traditional IR features.

The paper is organized as follows. Section 2 discusses the main contribution that of possible supervised approaches to match comments and segments. Section 3 proposes two novel feature representations: *enriched* representation using ESA [1] and coreference features. Section 4 contains the experimental results. The paper closes with conclusions and a discussion.

2. MATCHING COMMENTS WITH SEGMENTS

There is no readymade solution strategy for the problem statement given before. To this end we investigate both supervised and unsupervised strategies. To begin the discussion, we assume that we are given a dataset $\mathcal{D}$ defined on a collection of articles $\{A_1, \ldots, A_n\}$ where

$$\begin{aligned}\mathcal{D} &= \{ (c_{ij}, S_{ij}) \mid c_{ij} \in C(A_i), \\ & \quad S_{ij} = \{s | s \in S(A_i) \wedge c_{ij} \text{ is about } s\},\ i = 1, \ldots n\}\end{aligned} \tag{1}$$

2.1 Supervised Approach for Matching comments with Segments

In this section we formulate the supervised learning problem in two ways, namely the structured learning approach and the binary classification approach.

2.1.1 Structured Learning for locating segments

In the structured learning setup [4], one considers the problem of assigning labels $y \in \mathcal{Y}$, one of many labels, to an observation $x \in \mathcal{X}$. Structured classification proceeds by assuming that, given an x and w, the following function

$$f(x) = \arg\max_{y \in \mathcal{Y}} \mathbf{w}^\top \psi(x, y)$$

is tractable where $\psi(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is a suitably defined feature function.

In this paper, we discuss relevant details of structured learning as applied to the problem at hand. For a detailed review of structured learning see [4].

In the structured learning setup we make an assumption that in (1), each c_{ij} is mapped to *only one* segment, in other words $|S_{ij}| = 1$, where the cardinality of S_{ij} is denoted by $|S_{ij}|$. The resultant dataset is written as $\mathcal{D} = \{(c_{ij}, s_{ij})\}$, where s_{ij} denotes the segment correspond to c_{ij}. We wish to explore structured learning framework for this dataset.

We begin by assuming that $\psi(c, s) : C(A) \times S(A) \to \mathbb{R}^d$ is a given feature function. For more description of ψ, see Section 3.3. Given a $\mathbf{w} \in \mathbb{R}^d$, we wish to infer the segment associated with a comment c in a particular article A

$$f(c) = \arg\max_{s \in S(A)} \mathbf{w}^\top \psi(c, s) \tag{2}$$

which can be computed by straightforward enumeration as $|S(A)|$ is low.

The learning of $\mathbf{w}$ on a training set depends on the specific loss function. For the problem at hand we use

$$\Delta(s, f(c_{ij})) = \mathbf{1}_{s_{ij} \neq f(c_{ij})} \tag{3}$$

where $\mathbf{1}$ is the indicator function. The idea is to choose $\mathbf{w}$ such that for each comment $c_{ij} \in C(A_i)$, the score of the correct segment s_{ij} will be highest among all the other segments. This can be ensured by requiring that $\mathbf{w}^\top \psi(c_{ij}, s_{ij}) \geq \mathbf{w}^\top \psi(c, s), \quad \forall s \in S(A) - s_{ij}$. This immediately motivates the following optimization problem.

$$\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N} \sum_{i=1}^{n} \sum_{j=1}^{|C(A_i)|} \xi_{ij}$$

subject to

$$\begin{aligned}\mathbf{w}^T \psi(c_{ij}, s_{ij}) &\geq \mathbf{w}^\top \psi(c_{ij}, s) + 1 - \xi_{ij}, \\ & \qquad \forall s \in S(A_i) - s_{ij},\ c_{ij} \in C(A_i), \text{ and} \\ \xi_{ij} &\geq 0\end{aligned}$$

Lemma: *At optimality,*

$$\sum_{i=1}^{n} \sum_{j=1}^{|C(A_i)|} \xi_{ij} \geq \Delta(s_{ij}, f(c_{ij}))$$

where Δ is defined in (3) and f is defined in (2).

Proof:
Let $s = f(c_{ij}) \neq s_{ij}$ be the segment identified by (2). Then see that

$$\xi_{ij} \geq 1 - \left(w^\top \psi(c_{ij}, s_{ij}) - w^\top \psi(c_{ij}, s) \right)$$

By definition of s the difference of scores

$$\mathbf{w}^\top \delta_{ij} \psi(s) = \left(\mathbf{w}^\top \psi(c_{ij}, s_{ij}) - \mathbf{w}^\top \psi(c_{ij}, s) \right) \leq 0 \tag{4}$$

which implies that $\xi_{ij} \geq 1$. The claim holds whenever the assignment implemented by (2) disagrees with the correct segment s_{ij}. Noting that ξ_{ij} is lower-bounded by 0, the claim is proved. □

The optimization problem pertaining to the learning problem is a Quadratic Program (QP) involving d variables and $\sum_{i=1}^{n} |C(A_i)|(|S(A_i)|-1)$ constraints. The values of $|C(A_i)|$ are around 10 and that of $|S(A_i)|$ are often $10-15$ depending on the segmentation. The number of constraints is linear in the number of articles. In general solving such problems is hard as there could be large number of constraints. See [4] for the general case. Fortunately, we can solve the original problem directly by using a Convex QP as the number of constraints is small.

2.1.2 Binary Classification

In the previous section we discussed a strategy based on structured learning. But as noted the setting applies when dataset (1) is restricted to have only one segment assigned to each comment. We use the same $\psi(c, s)$ and define a classification problem. We assign a label $Y : C(A) \times S(A) \to \{\pm 1\}$ to all (comment, segment) pairs as follows: $Y(c_{ij}, s) = 1, \quad \forall s \in S_{ij}$. Using this one could immediately rewrite (1) as a dataset $D = \{(\phi(c_{ij}, s), Y(c_{ij}, s)\}$, which is essentially a binary classification problem. This immediately motivates a SVM-like approach

$$\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N} \sum_{i=1}^{n} \sum_{j=1}^{|C(A_i)|} \sum_{k=1}^{|S(A_i)|} \xi_{ijk}$$

subject to

$$\begin{aligned}Y(c_{ij}, s_{ik}) \mathbf{w}^\top \psi(c_{ij}, s_{ik}) &\geq 1 - \xi_{ijk}, \\ & \qquad \forall c_{ij} \in C(A_i), s_{ik} \in S(A_i), \text{ and} \\ \xi_{ijk} &\geq 0\end{aligned}$$

Again we use a generic SVM solver for this purpose.

2.2 Unsupervised Matching

One could choose to ignore the segments and instead choose to learn a similarity function by exploratory analysis. The key idea is to design a similarity function $sim(c, s) : C(A) \times S(A) \to \mathbb{R}$ such that for a fixed c it will enable us to evaluate how similar a comment c is to a particular segment s. To this end we consider that a comment or segment is represented by a vector as represented by $\phi(\cdot)$. Given such a representation, we define the the cosine similarity

$$sim(c, s) = \frac{\phi(c)^\top \phi(s)}{\|\phi(c)\| \|\phi(s)\|} \qquad \text{(COSINE)}$$

The vectors $\phi(c)$ and $\phi(s)$ are obtained from various feature functions discussed in Section 3.

3. REPRESENTATION

In our previous work [3] we have experimented with several feature representations for comment alignment namely bag of words (BOW), semi-supervised PLSA, and Latent Dirichlet Allocation (LDA). However because of extremely short length of comments and segments they are not very useful to capture the *semantics*. To address this issue we propose to leverage external corpus to have a more *enriched* representation. To this end we propose Explicit Semantic Analysis and coreference features.

3.1 Explicit Semantic Analysis

Unlike the above latent space approaches like LDA and SS-PLSA described above, explicit semantic analysis (ESA) uses a rich external corpus for representation [1]. Wikipedia is usually the corpus of choice. Let a be a Wikipedia article. For a term t, let $k_{t,a}$ denote the strength of association of t to a. Since Wikipedia has pages for most of the important concepts (people, places, events, organizations, etc.), we can consider $k_{t,a}$ as the affinity of t to the concept a and the vector $\mathbf{k}_t$ of all such weights as a representation of t in the concept space represented by Wikipedia.

Given a document d, the feature vector for d is given by

$$\phi(d) = \sum_{t \in d} w_{t,d} \mathbf{k}_t$$

In our case the document d could be a segment or a comment.

3.2 Coreference features

In addition to bag of words features, we also add a class of lexical and semantic features we call *coreference features* because of their widespread use in coreference resolution.

These features are listed below.

1. Number: Singular, Plural
2. Gender: Male, Female, Neuter, Unknown
3. Semantic class: Person, Location, Organization, Date, Time, Money, Percent, Object
4. Animacy: Human, Animal

These features are calculated using dictionaries, Stanford NER and Wordnet. Using this, we associate a 16-dimensional binary feature with each token t.

As mentioned before, these features are very popular in coreference resolution literature [2]. We can think of comment mapping as more complex form of reference resolution where we want to identify whether a comment directly or indirectly refers to an article segment. While these features are defined at token level for references in the coreference literature, we define them at segment or comment level. More formally, let f_t be the binary feature vector associated with a token t. The feature vector for a segment/comment s is given by

$$\sum_{t \in s} f_t$$

3.3 Cumulative Feature Vector

Given a comment c and a segment s, we can calculate features: BOW, LDA topics, SS-PLSA topics, Coreference features, and ESA for both c and s. For each of these features, we then calculate the cosine similarity. The final feature vector consists of BOW cosine similarity, Lucene score, Indri Score, LDA topic similarity, SS-PLSA similarity, Coreference feature similarity, and ESA similarity.

4. EXPERIMENTAL RESULTS

In this section we investigate several key questions related to the models proposed before. Broadly there are three questions one would like to explore.
(**a**) The effectiveness of the feature representations. We investigate empirically the effect of *enrichment*.
(**b**) Does supervision increase performance?
(**c**) Comparison between structured approach vis-a-vis binary classification.
In this section we explore these issues.

4.1 Experimental Setup

Datasets.

We created a dataset D by collecting 208 news articles along with ≈ 10 comments for each of the articles from `http://news.yahoo.com`. There were a total of 1079 comments in the corpus. We also created another dataset $D_{enriched}$ with same articles and comments as of D but with each of the articles enriched with additional (4–8) related articles. The related articles for an original article have been found by Google news search (`http://news.google.com`) with title of the original article as the search query.

The articles have an average length of 383.9 words (after stemming and stop word removal), 19.3 segments and 5.3 comments. The average length of comments is 24.6 words. In $D_{enriched}$, an average of 1316.2 words and 56.2 segments were added per article.

We created ground truth for for all the comments of all articles in the dataset. We have experimented with both the datasets D and $D_{enriched}$ with different methods discussed in Section 3.

Evaluation Metrics.

Let S_{ij} be the set of true related article-segments (found by human inspection) for comment c_{ij} as in (1).

If $|S_{ij}| > 1$, then c_{ij} has multiple related article-segments or if $|S_{ij}| = 0$, c_{ij} has no related article-segment.

Let r_{ij} be the retrieved result for comment c_{ij}. We consider this to be correct if $r_{ij} \in S_{ij}$.

The *Retrieval Index* is defined as:

$$RI = \frac{|\bigcup_{i=1}^{M} \{c_{ij} \in C(A_i) : r_{ij} \in S_{ij}\}|}{|\bigcup_{i=1}^{M} C(A_i)|}$$

i.e., RI is the ratio of number of correctly matched comments and total number of comments. Even if a comment is associated with multiple segments, we use only one segment from the classifier. It can be seen that most of the comments are short and are usually about one specific topic discussed in the article. Hence the above formulation of using only one retrieved segment is reasonable.

All the numbers reported here are the results of 10-fold cross validation.

Technique	RI
Unsupervised	58.9%
Binary SVM	62.6%
SVMStruct	**63.5%**

Table 1: RI for unsupervised and supevised matching techniques.

Features for supervised classification: To provide a rich feature space for the classification models, we use a binned representation of each feature. Let s be the value of a feature S (e.g., BOW similarity score). We choose a set T_S of thresholds for S. Then the binned representation of s is

$$< \mathbf{1}_{[s>t]} : \forall t \in T_S >$$

We thus convert each underlying feature to a T_S-dimensional feature capturing $T_S + 1$ bins. For example, if $s = 0.35$ and $T_S = \{0.1, 0.5, 0.9\}$, then the binned feature representation of s is $< 1\ 0\ 0 >$. In our experiments, $|T_S| = 10$ worked well.

4.2 Performance Comparison

Effectiveness of supervised techniques.

In the first experiment, for each comments and the segments of the article it corresponds to, we calculate the following seven features: BOW, LDA (or enriched LDA, denoted eLDA), SS-LDA, ESA, Coreference features, Lucene, and Indri. The features were binned as discussed above. The best results for SVMStruct were obtained with binned representation while that for binary SVM were obtained with non-binned representation. Since the number of matching segments for a comment is much smaller than the number of non-matching comments, the training data is unbalanced. We randomly sampled a subset of negative examples. The best C obtained using parameter sweeps for StructSVM is 0.5 while that for binary SVM is 1.1.

Table 1 shows RI for unsupervised matching as well as binary and structured SVM classifiers.

1. Supervised approaches have higher RI compared to the unsupervised approach. Binary SVM has a RI gain of 6.3% and structured SVM has a gain of 7.8% compared to unsupervised matching.

2. Structured classification performs better than binary classification by 1.4%.

Effectiveness of Enrichment.

The best performance using unsupervised matching is achieved by combining eLDA with BOW on enriched corpus – 57.8% of the comments are matched correctly (Table 2(b)). eLDA+BOW improves LDA by 124% and BOW by 3.4%.

It should be noted that all feature bring down the effectiveness of BOW when combined with it except eLDA. This shows that enriched topics provide a representation that is complementary to BOW.

It can also be seen that while ESA outperforms eLDA (by 14%), eLDA+BOW is more effective compared to ESA+BOW (by 8%).

Note: In Table 2, only eLDA features are computed on the enriched dataset as other features like BOW, PLSA, etc. cannot exploit the enriched dataset.

Feature	RI
BOW	55.9%
Lucene	57.2%
Indri	56.1%
SS-PLSA	51.4%
SS-PLSA+BOW	54.1%
ESA	41.0%
ESA + BOW	53.8%
LDA	25.8%
LDA+BOW	53.6%
Coreference	28.5%
Coreference + BOW	51.7%
Random	7.2%

(a) Performance on the Comment Dataset (D).

Feature	RI
eLDA	36.0%
eLDA+BOW	**57.8%**

(b) Performance on the Enriched Dataset ($D_{enriched}$).

Table 2: Effectiveness of enrichment using unsupervised matching. The table show RI scores for different features on two datasets – D and $D_{enriched}$.

5. CONCLUSIONS AND DISCUSSION

In this paper, we have proposed new approaches to the problem of aligning comments to relevant parts of the article to reduce the readers' cognitive burden. We first pose this problem as a structured classification problem as well as a binary classification problem. We show that structured classification outperforms binary classification. We then explore two additional features for this task: Explicit Semantic Analysis and coreference features. Explicit Semantic Representation using Wikipedia does not perform well. This implies corpus-specific enrichments are more effective compared to generic enrichment schemes. In future work, we plan to explore increasing the effectiveness of corpus-specific enrichment schemes. For enrichment, the source and selection of documents, number of documents needed, etc. are very important and we plan to come up with systematic guidelines for these.

6. REFERENCES

[1] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In *IJCAI*, 2007.

[2] A. Rahman and V. Ng. Narrowing the modeling gap: A cluster-ranking approach to coreference resolution. *JAIR*, 2011.

[3] D. K. Sil, S. H. Sengamedu, and C. Bhattacharyya. ReadAlong: Reading articles and comments together. In *WWW*, 2011.

[4] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Support vector learning for interdependent and structured output spaces. *JMLR*, 2005.