The Hong Kong Polytechnic University at DUC2005

Wenjie Li, Wei Li, Baoli Li, Qing Chen and Mingli Wu

Department of Computing
The Hong Kong Polytechnic University
{cswjli, cswli, csblli, csqchen, csmlwu}@comp.polyu.edu.hk

Abstract

This paper discusses the query-based multi-document summarization techniques implemented by the Hong Kong Polytechnic University at DUC 2005. The summarization system is built under the framework of MEAD. In addition to borrow the features provided by MEAD for text summarization, including centroid and sentence length etc., we also introduce the entity-based, pattern-based, term-based and semantic-based features in particular for query relevance judgment. This is our first time to participate in DUC. However, the evaluation results are encouraging. Our system ranks competitively in DUC 2005, especially in ROUGE evaluations.

1 Introduction

The task of DUC 2005 is query-based multi-document summarization (QMS), which requires creating from a set of relevant documents a brief, well-organized and fluent summary to the need for information that cannot be met by just stating a name, date or quantity. QMS is an emerging research area coupling Multi-document Summarization and Question and Answering (QA) techniques. From Multi-document Summarization perspective, QMS needs to generate an information-complete and coherent summary. From QA perspective, a summary should fulfill users' information need, i.e. the summary should not only be relevant to the queries, but also provide answers to the queries in the form of natural language questions. These render QMS task complicated.

To ensure the fast development within the time limitation, we build our system under the framework of MEAD[1]. MEAD is developed by the University of Michigan and got competitive performance in DUC 2004 [1]. Most importantly, it provides a good extractive-based summarization framework for reuse. It allows integrating any additional user-designed features and can automatically join features together to evaluate the importance of a sentence to be included in a summary.

The free-downloaded version of MEAD provides three basic types of features for multi-document summarization. They are centroid, sentence position and sentence length. Centroid features [2] which extract the thematic information help the system measure the importance of a sentence within a document set. Sentence length features cut off too long or too short sentences. These two types of features are reserved in our system. However, sentence positions are not taken into account. The assumption that the most important information is conveyed in the beginning or the end of a document may not accommodate to the requirement of fulfilling users' information need.

The MEAD-based features mentioned above measure the importance of a sentence from the summarization perspective. We then design proper features from QA perspective with more emphasis on how to retrieve potential answers to the questions in the query. In addition to the conventional use of term-based[2] features (term matching overlapping, indicated by F_T), the other three types of features: entity-based features (named entity matching, indicated by F_E), pattern-based features (indicated by F_P) and semantic-based features (indicated by F_S) are also investigated.

Named entities are of particular importance in serving for the agents, the patients, the times and locations of events. They are also used to describe objects and the associated attributes, e.g. a person and his age, an organization and its abbreviation, a location and its alias. Entity-based features can be useful in retrieving the objects and events requested by users. As a matter of fact, the contribution of named entities has been quite significant in QA [3] and summarization [4]. The use of linguistic patterns for answer extraction based on answer predication has also been proven to be a very effective strategy [5], especially in answering definition

[1] Free downloadable from http://www.summarization.com/mead/.

[2] A term can be either a word or a named entity.

question [6]. Previous observation shows definition and description questions take a large proportion of DUC 2005 query set. Thus, pattern-based features are integrated in the system. Moreover, we consider making use of semantic information to excavate the sentences which are semantically relevant to queries. To discover the semantic similarity between the words, word senses are disambiguated with WordNet[3] [7].

Besides, our summarization system performs a series of preprocessing including sentence segmentation, word stemming and named entity recognition, and post-processing including sentence compression and re-ordering. It finally got very competitive performance in DUC 2005, especially in ROUGE evaluations. After the evaluation, we further experiment on different combinations of the features in order to reveal their importance. The experimental results tell that the pattern-based features outperform the entity-based features. And the liner combination of the four features designed by ourselves performs better than basic MEAD-based features.

The rest of this paper is organized as follows. Section 2 introduces the overview of the system. Feature design is detailed in Section 3. In Section 4, the experiments and evaluation results are presented. Finally, the last section discusses the future work and concludes the paper.

2 System Overview

Given a query and a document set, the summary is generated in four steps, as shown in Figure 1, (1) pre-processing (2) feature extraction (3) sentence evaluation and (4) post-processing.

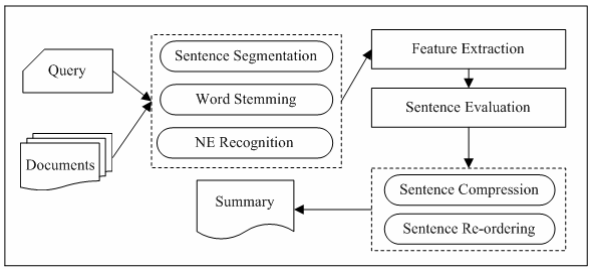


Figure 1 System overview

2.1 Pre-Processing

In preparatory step, the sentences are first segmented, the words are stemmed and the named entities are tagged. Word stemming is necessary because the words in different forms but with the same root, e.g. crime and crimes, often share the same meaning. Pre-processing is carried out with a collection of tools, sentence segmentor[4], Porter Stemmer[5] and GATE[6] . We choose to use four types of named entities provided by GATE, i.e. <Person>, <Organization>, <Location> and <Date>, considering they can facilitate the task. In addition, we add the type of <Number>, which can be recognized by the following regular expression:

Number: ([0-9]|,|.)*([a-z]|%)

It identifies the real numbers (e.g. 15, 1.5), the percentages (5%) and the distances (5m). Altogether, five types of named entities are involved in potential answer extraction and answer validation.

2.2 Feature Extraction

As introduced in Section 1, five types of features are extracted from queries and document sets. We refer to them as MEAD-based (F_M), term-based (F_T), entity-based (F_E), pattern-based (F_P), and semantic-based features (F_S). The explanation will be detailed in Section 3.

2.3 Sentence Evaluation

MEAD provides an extractive-based summarization framework. It calculates the score for each sentence based on the proposed features. The top ranked sentences are selected as the summary. Assume W_M, W_T, W_E, W_P and W_S are the weights of the features F_M, F_T, F_E, F_P and F_S, respectively. The *score* of the sentence s is calculated as,

$$Score(s) = \begin{cases} W_M\alpha_{F_M} + W_T\alpha_{F_T} + W_E\alpha_{F_E} + W_P\alpha_{F_P} + W_S\alpha_{F_S} ; & \alpha_{F_L} \in [\theta_1,\theta_2] \\ 0 ; & \alpha_{F_L} \notin [\theta_1,\theta_2] \end{cases}$$

where α is the functions assigning sentence scores by each individual features. F_M combines centroid and QueryCosineNoIDF, a query-relevant feature, and α_{F_M} is provided by MEAD; Sentence length, is indicated by F_L, α_{F_L} is the sentence length. $[\theta_1,\theta_2]$ limits the maximum and maximum length. The rest four score calculation methods will be described in Section 3. In the system, weights and length limitations are empirically determined as the following,

$W_M : W_T : W_E : W_P : W_S = 1:2:2:2:1$, θ_1=9, θ_2=50.

2.4 Post-Processing

Since the summary is limited to 250 words, we remove some unnecessary text segments in order to allow more relevant information included into the summary. A set of heuristic rules are devised for this purpose:

[c1] If the segment occurs as "--" segment "--" or "(" segment ")", the segment is removed.

[c2] If the segment length is less than 5 words, and

[3] Free downloadable from http://wordnet.princeton.edu/.

[4] Free downloadable from http://l2r.cs.uiuc.edu/~cogcomp/tools.php.

[5] Free downloadable from http://www.tartarus.org/~martin/PorterStemmer.

[6] Free downloadable from http://gate.ac.uk/.

a) the segment includes an entity tag <Person> and a word derived from "say", e.g. said, it is removed; or
b) the segment includes a word derived from "say" and its length is less than 3 words, it is removed.

For the task of multi-document summarization, sentence re-ordering is necessary to ensure the coherence of the summary. The following two criteria are applied:

[r1] If the two sentences are selected from the different documents, they are ordered according to document published dates.
[r2] If the two sentences come from the same document, their order remains the same as it is in the document.

3 Feature Design

The query Q provided by DUC 2005 includes a title (T), a narrative (N) and a Granularity (G), i.e. $Q=\{T, N, G\}$. The title T is a sequence of terms (t), i.e. $T=\{t_i, i=1,\ldots,m\}$. A term can be a word (w) or a named entity (ne). The narrative N is a set of sub-queries (q), i.e. $N=\{q_i, i=1, \ldots, u\}$. The sub-query q or the sentence s can be represented by a sequence of terms as well, i.e. $q_i=\{t_i, i=1, \ldots, v\}$ and $s_i=\{t_i, i=1, \ldots, w\}$. The granularity G of the query Q can be specific or general. The distinction of query granularity is not under our consideration.

3.1 Term-based Feature

This is the basic feature widely used in information retrieval. It measures the relevance of the sentence to the given query based on the number of the words appearing in them. The assumption is that if a word or a named entity occurs in both the sentence and the query, it should contribute to their relevance. Stop-words (e.g. "the", "have" in [q1]), interrogative words (e.g. "who" in [q1]) or the first words in queries (e.g. "name" in [q2]) are excluded. A list of 200 words is used to filter stop-words.

[q1] Who has criticized the World Bank and… [d331f]
[q2] Name the countries involved. [d301i]

Given a sentence s and a query Q, the score of the term-based feature $\alpha_{F_w}(s,Q)$ is calculated as:

$$\alpha_{F_T}(s,Q)=\frac{1}{Z}(\lambda_1\alpha_{F_T}(s,T)+\lambda_2\alpha_{F_T}(s,N))$$
$$=\frac{1}{Z}(\lambda_1\alpha_{F_T}(s,T)+\lambda_2\sum_{i=1}^{i=|N|}\alpha_{F_T}(s,q_i))$$
$$=\frac{1}{Z}(\lambda_1\sum_{i=1}^{i=|s|}\sum_{j=1}^{j=|T|}I(t_i,t_j)+\lambda_2\sum_{i=1}^{i=|N|}\sum_{j=1}^{j=|s|}\sum_{k=1}^{k=|q_i|}I(t_i^s,t_k^{q_i}))$$

where $I(t_i,t_j)=\begin{cases}1 & t_i=t_j \\ 0 & t_i\neq t_j\end{cases}$ and

$$Z=\max(\lambda_1\alpha_{F_T}(s,T)+\lambda_2\alpha_{F_T}(s,N))$$
$$=\lambda_1|s||T|+\lambda_2|s||N|\sum_{i=1}^{i=|N|}|q_i|$$

$|N|$ is the number of the sub-queries in the narrative N; $|s|$ and $|q_i|$ are the number of terms in s and q_i respectively. Z is a normalization factor. λ_1 and λ_2 are weights associated to the title and narrative respectively.

Intuitively, query title should be more informative than query narrative. However, the experiments illustrated in Section 4 show us the opposite results. In order to avoid duplicating calculation, the terms occurring in both title and narrative are excluded from narrative.

3.2 Entity-based Feature

We distinct the named entities in the sub-queries as two categories, either the term entities, which are tagged by GATE, e.g. "World Bank" is tagged with <Organization> in [q1], or the Question Type (QT) entities, e.g. "who" in [q1] implies the <Person> entity. Question type indicates what information the question is looking for. Both of these two categories of named entity are useful for summary sentence extraction. Moreover, QT can especially help validate the appropriateness of the sentences to be included in a summary. For example, if a sentence contains the named entity tagged as <Person>, it will be most likely to be an answer for question [q1]. However, a question does not always ask for a named entity. For the question with a non-entity question type, the validation process is not applicable.

We describe how to determine the question type now. For the sub-queries beginning with the interrogative words "who", "where" and "when", a straightforward mapping between these interrogative words and the types of the named entity to be questioned is constructed, e.g. "who"-<person>, "where"-<location> and "when"-<Date>. In contrast, the sub-queries beginning with the word "Name" as in [e2] and the interrogative words "which" and "what" need a more complex deduction. We consider the four patterns, "which + noun", "what + noun", "what + be + noun", and "name + noun", take the words "which", "what" and "name" as indicators and then follow the following four steps to deduce the question type:

[ne1]. Find the indicator in the query;
[ne2]. Fetch the noun[7] after the indicator;
[ne3]. Find the hypernyms of the noun in WordNet.
[ne4]. Determine the question type.
a) If one of the words provided by step [ne3] is "person", "organization", "location", "date" or "number", then the question type is the same as

[7] We use the free trail version of CLAWS as our part-of-speech tagger: http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/trial.html .

it. For example, if the word "location" is a hypernym of the word "country", the question type of [q2] is <location>.

b) If more than one named entity is found, the closest one in the hypernyms is selected.

Assume the query Q and the sentence s contain a set of entities $\{e_i^Q, i=1, \ldots, m\}$ and $\{e_j^s, j=1, \ldots, u\}$. The corresponding tag sets are $\{tag_i^Q, i=1, \ldots, n\}$ $(n \le m)$ and $\{tag_i^s, j=1, \ldots, v\}$ $(v \le u)$, respectively. The entity-based features are calculated by

$$\alpha_{F_E} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{v} I(tag_i^Q, tag_j^s); \quad I(tag_i^Q, tag_j^s) = \begin{cases} 1 & tag_i^Q = tag_j^s \\ 0 & tag_i^Q \neq tag_j^s \end{cases}$$

3.3 Pattern-based Feature

Both the term-based and the entity-based features neglect the orders of matching. This may sometimes lead to the mistakes like including a sentence with the phrase "World Bank criticized <Person>" into the summary responding to [q1]. Even thought such a phrase contains all the terms and the named entity of the same type as the question, it is irrelevant to the expected response to [q1]. We hence decide to give extra bonus to the sentences which matches the term orders in the sub-queries. Considering that a query is generally longer than the matched sentence segments, we decompose the whole sub-query into pieces of shorter segments. Three types of patterns are then extracted automatically.

[p1]. (entity tag, entity tag) or (entity tag, entity tag)
The question type (if *QT* exists) and a named entity tag or two named entity tags are combined into a pattern. They can be separated by the other terms.

[p2]. (entity, entity tag) or (entity tag, entity)
A named entity is combined with the tag of another named entity or *QT* into a pattern. They can be separated by the other terms.

[p3]. (word, entity tag) or (entity tag, word)
A word is combined with the tag of another named entity or *QT* into a pattern. The word must be a non-stop word adjacent to that named entity.

The original order between the words and named entities in sub-queries are reserved in the patterns. After pattern extraction, the duplicated patterns are removed. Here is an illustrative example.

> *Question:* Who has criticized the World Bank? [d331f]
>
> *Step* 1*:* named entity and *QT* recognition
> <Person> has criticized the <Organization>?
>
> *Step* 2*:* pattern extraction
> (<Person>, <Organization>); (<Person>, <World Bank>); (<Person>, criticize); (criticize, <Organization>).

The pattern-based feature is calculated as a binary feature. If a query pattern occurs in a sentence, the feature value α_{F_P} is set to 1, otherwise α_{F_P} is 0.

3.4 Semantic-based Feature

The sentences with different but semantic related words receive lower score with term-based features, which in turn may result in being ignored in the summary. To avoid this problem, we calculate the semantic overlapping between the query title T and sentence s with the following equation.

$$\alpha_{F_S}(s,T) = \frac{1}{Z}\sum_{i=1}^{i=|T|} \max_{j=1}^{j=|s|} I(t_i, t_j)$$

$$I(t_i, t_j) = sim(t_i, t_j), \quad (t=w)$$

$$Z = \max \sum_{i=1}^{i=|T|} \max_{j=1}^{j=|s|} I(t_i, t_j)$$

$$= \sum_{i=1}^{i=|T|} I(t_i, t_i)$$

where $sim(t_i, t_j)$ is *lesk*-similarity [8] based on the word senses. Z is a normalization factor. WordNet-Similarity-0.15[8] is used to calculate *lesk*-similarity. It is quite common for a word with more than one sense. Senses are disambiguated with the package WordNet-SenseRelate-AllWords, which is able to determine word senses in a given context [10]. However, disambiguation on all the document sets is a time-consuming process. In the official run, we simply choose to use "sense1" disambiguation scheme which does not take into consideration the context, and assumes that the correct sense for a word is its dominant one in WordNet [7]. Taking "sense1" disambiguation scheme seems to be a good compromise between efficiency and accuracy.

The calculation between the query narrative N and the sentence s can be derived in a similar way. In our official run, however, we only consider the query's title, because we believe that the title of a query is a highly condensed summary of its narrative, and the latter may contain much noise.

4 Evaluation

4.1 Data Set and Evaluation Metric

DUC 2005 provides fifty document sets for evaluation. Each document set includes 25~50 documents and an associated query. Each query has a query title, a query narrative including a set of sub-queries and a granularity. All submitted systems are either manually or automatically evaluated according to the summary's linguistic quality, its responseness, overlapping with human generated summary (ROUGE-2, ROUGE-SU4) and Pyramid. Besides these evaluations, we made post-

[8] Free downloadable from http://search.cpan.org/dist/WordNet-Similarity/.

experiments on different combinations of the features in order to explore the importance of the proposed features.

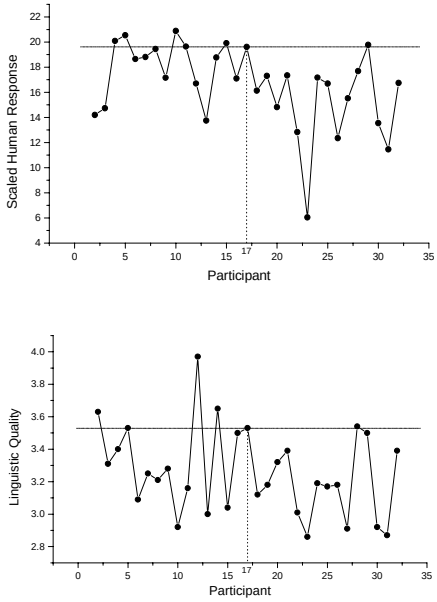
4.2 Evaluation Results

4.2.1 Results Provided by DUC2005

Among the 31 submitted systems, our system ranks the 7th in responseness evaluation, the 6th in linguistic quality evaluation, and the 2nd in ROUGE evaluations (both ROUGE-2 and ROUGE-SU4). The system also got the 6th in Pyramid evaluation and 4th the modified Pyramid evaluation out of 25 systems (the official score from "processed_pans.txt"), as shown in Figure 2. It is natural that the ROUGE evaluation results are better than the human evaluated results since we do not make any effort on co-reference resolution.

4.2.2 Experiments on Features

We investigate on the impacts of different combinations of the features, and present the ROUGE evaluation results in Table 1. It is interesting to see that the entity-based features take in negative impacts on the result when it combines with the term-based features. However, when combined with both the term-based and the pattern-based features, it does show a degree of contribution in ROUGE-2. There are two possible reasons. Firstly, entity-based features are not independent to other features and cannot be used independently. Secondly, the five types of named entities recognized by GATE are not sufficient to define the query types. It is also shown that F_T outperforms F_M and F_S significantly and the combination all the features produce the best result.



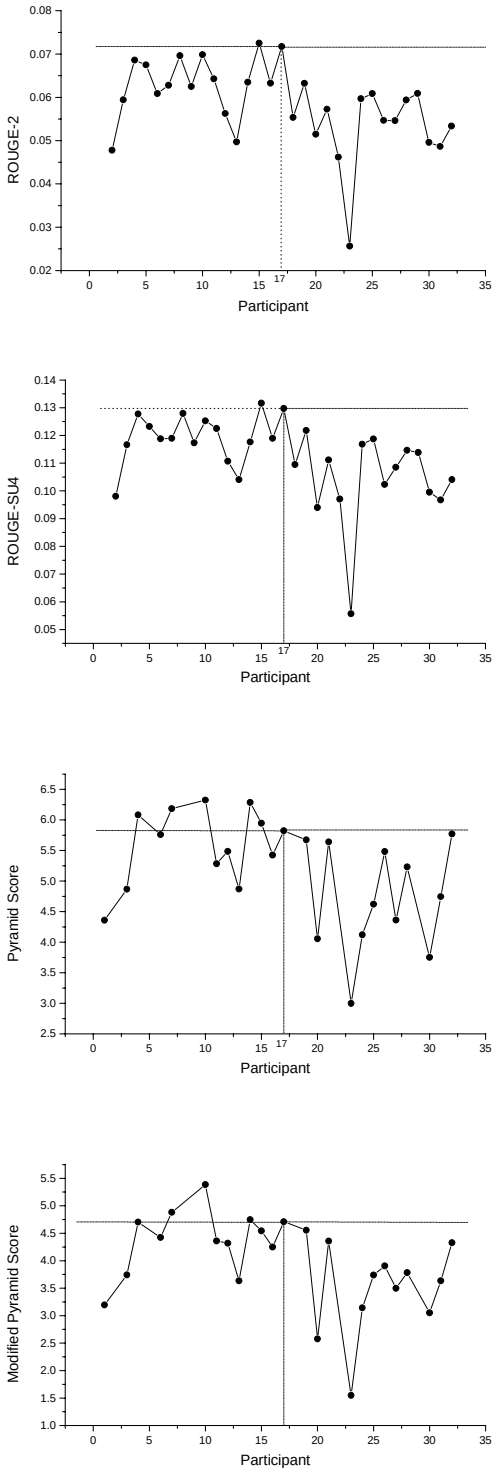


Figure 2 Experiment Results provided by DUC 2005

Table 1 Experimental results of the different feature combinations

Features	ROUGE-2	ROUGE-SU4
F_M+2*F_T+2*F_E+2*F_P+F_S	0.07174	0.12972
F_T+F_E+F_P	0.07117	0.12773
F_T+F_P	0.07054	0.12843
F_T+F_E	0.06844	0.12654
F_T	0.06878	0.12658
F_M	0.06447	0.12295
F_S	0.06053	0.11572

As mentioned in Section 3, we wonder whether the query titles should be assigned with the higher or the same weights with the query narratives. Experiments are conducted to justify the hypothesis.

Table 2 Experimental results of the different weights assigned to the query titles and the narratives

Features	ROUGE-2	ROUGE-SU4
$\lambda_1 : \lambda_2 = 1:1$	0.06878	0.12658
$\lambda_1 : \lambda_2 = 2:1$	0.06732	0.12458

The experimental results shown in Table 2 seem to suggest that it is not necessary to stress the importance of the query titles. Query titles are often composed of abstract words (e.g. crime, project), which may contribute less during term matching.

8. Conclusion and Future Work

This is our first time to participate in DUC. Our query-based multi-document summarization system is built under the MEAD framework by integrating additional features. They are term-based, entity-based, pattern-based and semantic-based features.

Although our system has got competitive results, there are lots of rooms for improvement. An appropriate and wide-coverage named entity recognizer will improve the performance of the entity-based features. Introducing in-depth word sense disambiguation will make the semantic-based features more capable. Since the experiments show that the pattern-based features are very useful, we'd like to concentrate more on learning various patterns automatically. Moreover, we would like to study on the inter-relation between the sub-queries and intra-relation between the chunks within sub-queries to enhance query processing in the future.

References

[1].Erkan G. and Radev D. The University of Michigan at DUC 2004. In Proceedings of the Document Understanding Workshop. Boston, USA, May 6-7, 2004.

[2].Radev D., Jing H., Stys M., and Tam D. Centroid-based summarization of multiple documents. Information Processing and Management, 40:919-938, December 2004.

[3].Li X., Roth D., and Small K.: The Role of Semantic Information in Learning Question Classifiers. Proceedings of the International Joint Conference on Natural Language Processing, 2004.

[4].Regina Barzilay, Mirella Lapata. Modeling local coherence: an entity-based approach. In Proceedings of the 43rd Annual Meeting of the Association for. Computational Linguistics (ACL'05), pages 141–148, University of Michigan, 2005.

[5].Ravichandran, D., & Hovy, E. Learning surface text patterns for a question answering system. In Proceedings of the 40th Annual Meeting of the ACL, pages 41-47, 2002.

[6].Wesley Hildebrandt, Boris Katz, and Jimmy Lin. Answering Definition Questions with Multiple Knowledge Sources. Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004), May 2004, Boston, Massachusetts.

[7].Christiane Fellbaum. 1998. WordNet: an Electronic Lexical Database. MIT Press.

[8].Banerjee S. and Pedersen T. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In Proceedings of the Fourth International Conference on Computational Linguistics and Intelligent Text Processing (CICLING-02). Mexico City.

[9].Radev D., Blitzer J., Winkel A, Topper M., Celebi A., and Lam W.. MEAD Documentation. http://www.summarization.com/mead/.

[10]. Pedersen T., Banerjee S., and Patwardhan S. 2005. Maximizing Semantic Relatedness to Perform Word Sense Disambiguation, University of Minnesota, Supercomputing Institute, Research Report UMSI 2005/25.