MEAD - a platform for multidocument multilingual text summarization

Dragomir Radev[1], Timothy Allison[1], Sasha Blair-Goldensohn[2], John Blitzer[3], Arda Çelebi[4], Stanko Dimitrov[1], Elliott Drabek[5], Ali Hakim[1], Wai Lam[6], Danyu Liu[7], Jahna Otterbacher[1], Hong Qi[1], Horacio Saggion[8], Simone Teufel[9], Michael Topper[1], Adam Winkel[1], Zhu Zhang[1]

[1]University of Michigan, [2]Columbia University, [3]University of Pennsylvania, [4]USC/ISI, [5]Johns Hopkins University, [6]Chinese University of Hong Kong, [7]University of Alabama, [8]University of Sheffield, [9]University of Cambridge
radev@umich.edu

Abstract

This paper describes the functionality of MEAD, a comprehensive, public domain, open source, multidocument multilingual summarization environment that has been thus far downloaded by more than 500 organizations. MEAD has been used in a variety of summarization applications ranging from summarization for mobile devices to Web page summarization within a search engine and to novelty detection.

1. Introduction

MEAD is the most elaborate publicly available platform for multi-lingual summarization and evaluation. Its source and documentation can be downloaded from `http://www.summarization.com/mead`. The platform implements multiple summarization algorithms (at arbitrary compression rates) such as position-based, centroid-based, largest common subsequence, and keywords. The methods for evaluating the quality of the summaries are both intrinsic (such as percent agreement, cosine similarity, and relative utility) and extrinsic (document rank for information retrieval).

MEAD implements a battery of summarization algorithms, including baselines (lead-based and random) as well as centroid-based and query-based methods. Its flexible architecture makes it possible to implement arbitrary algorithms in a standardized framework. Support is provided for trainable summarization (using Decision trees, Support Vector Machines or Maximum Entropy). Finally, MEAD has been used in numerous applications, ranging from summarization for mobile devices to Web page summarization within a search engine and to novelty detection.

2. Architecture

MEAD's architecture consists of four stages. First, documents in a cluster are converted to MEAD's internal (XML-based) format. Second, given a configuration file (.meadrc) or command-line options, a number of features are extracted for each sentence of the cluster. Third, these features are combined into a composite score for each sentence. Fourth, these scores can be further refined after considering possible cross-sentence dependencies (e.g., repeated sentences, chronological ordering, source preferences, etc.) In addition to a number of command-line utilities, MEAD provides a Perl API which lets external programs access its internal libraries. A sample .meadrc file is shown in Figure 1.

All data in MEAD is stored as XML. The following DTDs are part of MEAD:

- cluster: a description of all related documents that will be summarized together,

```
compression_basis sentences
compression_absolute 1
classifier
/clair4/projects/mead307/source/mead/bin/default-classifier.pl
    Centroid 3.0 Position 1.0 Length 15 SimWithFirst 2.0
reranker /clair4/projects/mead307/source/mead/bin/default-reranker.pl
    MEAD-cosine 0.9 enidf
```

Figure 1: Sample .meadrc file. Using this configuration file, MEAD will produce a one-sentence summary using a linear combination of three features as the scoring function. From any sentence pair where the IDF-modified cosine similarity is higher than 0.9, one of the sentences will be dropped.

- docjudge: relevance judgements associated with the document or summary and a particular query and retrieval method,
- docpos: a part-of-speech annotated version of the document,
- docsent: a document, split into sentences,
- document: the raw document,
- extract: alisting of all sentence that should be in the summary,
- mead-config: MEAD's configuration parameters,
- query: a TREC-style query converted to XML,
- reranker-info: parameters for the rerankers,
- sentalign: a sentence-to-sentence alignment across languages,
- sentfeature: a list of feature values for a given document and feature names,
- sentjudge: manually annotated sentences for relevance within a cluster,
- sentrel: CST-style sentence-to-sentence relationships.

A few sample files conforming to these DTDs are shown in the Appendix.

3. Features

The following features are provided with MEAD. They are all computed on a sentence-by-sentence basis.

- Centroid: cosine overlap with the centroid vector of the cluster (Radev et al., 2004),
- SimWithFirst: cosine overlap with the first sentence in the document (or with the title, if it exists),
- Length: 1 if the length of the sentence is above a given threshold and 0 otherwise,
- RealLength: the length of the sentence in words,
- Position: the position of the sentence in the document,
- QueryOverlap: cosine overlap with a query sentence or phrase,
- KeyWordMatch: full match from a list of keywords,
- LexPageRank: eigenvector centrality of the sentence on the lexical connectivity matrix with a defined threshold.

4. Classifiers

Four classifiers come with MEAD.

- Default: provides a linear combination of all features except for "Length" which is treated as a cutoff feature (see previous section),
- Lead-based: a baseline classifier that favors sentences that appear earlier in the cluster, as defined by the order of documents in the definition of the cluster,
- Random: a baseline classifier that extracts sentences at random from the cluster,
- Decision-tree: a machine learning algorithm, based on Weka (Witten and Frank, 2000) and trained on an annotated summary corpus.

5. Rerankers

The following rerankers are included in MEAD.

- Identity: this reranker does nothing; it preserves the scores of all sentences as computed by the classifier,
- Default: keep all scores, but skip sentences that are too similar (cosine similarity above a specific threshold) to sentence already included in the summary,
- Time-based: penalize earlier (or later, depending on the argument) sentences,
- Source-based: penalize sentences that come from particular sources,
- CST-based: this reranker applies different policies as determined by the cross-document structure of the cluster (Radev, 2000; Zhang et al., 2002),
- Maximal Marginal Relevance (MMR): this reranker is based on the MMR principle as formulated in (Carbonell and Goldstein, 1998).

6. Evaluation methods

The MEAD evaluation toolkit (MEADEval), previously available as a separate piece of software, has been merged into MEAD as of version 3.07. This toolkit allows evaluation of human-human, human-computer, and computer-computer agreement. MEADEval currently supports two general classes of evaluation metrics: co-selection and content-based metrics. Co-selection metrics include precision, recall, Kappa, and Relative Utility, a more flexible cousin of Kappa. MEAD's content-based metrics are cosine (which uses TF*IDF), simple cosine (which doesn't), and unigram- and bigram-overlap. An additional metric, relevance correlation, is available as an addon.

- Precision/recall: which sentences in the summary match the sentences in the human model,
- Kappa: takes into account interjudge agreement as well as the difficulty of the problem,
- Relative utility: similar to Kappa but allows for non-binary judgements in the model,
- Relevance correlation: there are two versions of this metric: Spearman (rank correlation) and Pearson (linear correlation); given a query, a search engine, and a document collection, Relevance correlation is high if a ranked list of the full documents in the collection given the query is highly correlated with a similar rankings based on the summaries of the documents.
- Cosine: cosine similarity against a human summary (or a set of human summaries),
- Longest-common subsequence: same as Cosine, but using the longest-common subsequence similarity measure,
- Word overlap: same as Cosine, but based on the number of words in common between the automatic and manual summaries,
- BLEU: based on the precision-oriented n-gram matcher developed by (Papineni et al., 2002).

7. Corpora

- SummBank: this is a large corpus for summary evaluation. It CD-ROM contains 40 news clusters in English and Chinese, 360 multi-document, human-written non-extractive summaries, and nearly 2 million single document and multi-document extracts created by automatic and manual methods. The collection was prepared as part of the 2001 Johns Hopkins summer workshop on Text Summarization (Radev et al., 2002).
- CSTBank: a smaller corpus, manually annotated at the University of Michigan for CST (Cross-document Structure Theory) relationships. CST relationships include subsumption, identity, fulfillment, paraphrase, elaboration/refinement, etc.

8. Utilities

The following utilities are included in MEAD:

- DUC conversion: scripts to convert DUC 2002–2004 style SGML documents into the MEAD format,
- Sentjudge to manual summary conversion: scripts to generate manual summaries from manual sentence-based non-binary relevance judgements,
- CIDR: a document clustering utility partially built over the MEAD API,
- Preprocessors: tools to convert plain text and HTML documents to the MEAD format.
- Sentrel utilities: tools to manipulate CST-style sentence relevance judgements.

9. Applications

MEAD has been successfully used in the following tasks: evaluate an existing summarizer, test a summarization feature, test a new evaluation metric, test a short-query machine translation system. It has also been used in major evaluations such as DUC (Radev et al., 2001a; Otterbacher et al., 2002; Radev et al., 2003) (text summarization) and TREC (question answering and novelty detection). Several systems have been built on top of MEAD, specifically NewsInEssence(Radev et al., 2001c; Radev et al., 2001b) (online news tracking and summarization), WebInEssence (Radev et al., 2001d) (clustering and summarization of Web hits), and WAPMead (in progress) (wireless access to summarization for email access).

10. History

MEAD v1.0 and v2.0 were developed at the University of Michigan in 2000 and early 2001. MEAD v3.01 – v3.06 were written in the summer of 2001 at Johns Hopkins University. As of Version 3.07, MEAD has been back to Michigan. The current release, 3.07, includes support for English and Chinese in a UNIX (Linux/Solaris/Cygwin) environment. Adding new (human) languages should be equally easy.

11. Acknowledgments

This work was partially supported by the National Science Foundation's Information Technology Research program (ITR) under grant IIS-0082884. All opinions, findings, conclusions and recommendations in any material resulting from this workshop are those of the participants, and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank a number of individuals for making MEAD possible: Fred Jelinek, Sanjeev Khudanpur, Laura Graham, Naomi Daniel, Anna Osepayshvili, and many others.

Appendix. Sample XML files

The following figures: 2, 3, 4, 5, 6, and 7 give illustrations of various XML files used by MEAD.

```
<?xml version='1.0'?>

<SENT-FEATURE>
<S DID="87" SNO="1" >
<FEATURE N="Centroid" V="0.2749" />
</S>
<S DID="87" SNO="2" >
<FEATURE N="Centroid" V="0.8288" />
</S>
<S DID="81" SNO="1" >
<FEATURE N="Centroid" V="0.1538" />
</S>
<S DID="81" SNO="2" >
<FEATURE N="Centroid" V="1.0000" />
</S>
<S DID="41" SNO="1" >
<FEATURE N="Centroid" V="0.1539" />
</S>
<S DID="41" SNO="2" >
<FEATURE N="Centroid" V="0.9820" />
</S>
</SENT-FEATURE>
```

Figure 2: Sentfeature object

```
<?xml version='1.0'?>
<SENT-JUDGE QID='551'>
<S DID='D-19980731_003.e' PAR='1' RSNT='1' SNO='1'>
                <JUDGE N='smith' UTIL='10'/>
                <JUDGE N='huang' UTIL='10'/>
                <JUDGE N='moorthy' UTIL='6'/>
</S>
<S DID='D-19980731_003.e' PAR='2' RSNT='1' SNO='2'>
                <JUDGE N='smith' UTIL='6'/>
                <JUDGE N='huang' UTIL='10'/>
                <JUDGE N='moorthy' UTIL='10'/>
</S>
<S DID='D-19980731_003.e' PAR='3' RSNT='1' SNO='3'>
                <JUDGE N='smith' UTIL='6'/>
                <JUDGE N='huang' UTIL='9'/>
                <JUDGE N='moorthy' UTIL='10'/>
</S>
<S DID='D-19981105_011.e' PAR='5' RSNT='2' SNO='7'>
                <JUDGE N='smith' UTIL='2'/>
                <JUDGE N='huang' UTIL='1'/>
                <JUDGE N='moorthy' UTIL='4'/>
</S>
</SENT-JUDGE>
```

Figure 3: Sentjudge object

```
<?xml version='1.0'?>
<!DOCTYPE DOC-JUDGE SYSTEM '/clair4/mead/dtd/docjudge.dtd'>

<DOC-JUDGE QID='Q-2-E' SYSTEM='SMART' LANG='ENG'>
    <D DID='D-19981007_018.e' RANK='1' SCORE='9.0000' />
    <D DID='D-19980925_013.e' RANK='2' SCORE='8.0000' />
    <D DID='D-20000308_013.e' RANK='3' SCORE='7.0000' />
    <D DID='D-19990517_005.e' RANK='4' SCORE='6.0000' />
    <D DID='D-19981017_015.e' RANK='4' SCORE='6.0000' />
    <D DID='D-19990107_019.e' RANK='12' SCORE='5.0000' />
    <D DID='D-19990713_010.e' RANK='12' SCORE='5.0000' />
    <D DID='D-19991207_006.e' RANK='12' SCORE='5.0000' />
    <D DID='D-19990913_001.e' RANK='20' SCORE='4.0000' />
    <D DID='D-19980609_005.e' RANK='20' SCORE='4.0000' />
    <D DID='D-19990825_018.e' RANK='1962' SCORE='0.0000' />
    <D DID='D-19990924_047.e' RANK='1962' SCORE='0.0000' />
</DOC-JUDGE>
```

Figure 5: Docjudge object

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE EXTRACT SYSTEM '/clair/tools/mead/dtd/extract.dtd'>

<EXTRACT QID='GA3' LANG='ENG' COMPRESSION='7'
SYSTEM='MEADORIG' RUN='Sun Oct 13 11:01:19 2002'>
<S ORDER='1' DID='41' SNO='2' />
<S ORDER='2' DID='41' SNO='3' />
<S ORDER='3' DID='41' SNO='11' />
<S ORDER='4' DID='81' SNO='3' />
<S ORDER='5' DID='81' SNO='7' />
<S ORDER='6' DID='87' SNO='2' />
<S ORDER='7' DID='87' SNO='3' />
</EXTRACT>
```

Figure 7: Extract Object

```
<?xml version='1.0'?>
<!DOCTYPE QUERY SYSTEM "/clair4/mead/dtd/query.dtd" >

<QUERY QID="Q-551-E" QNO="551" TRANSLATED="NO">
<TITLE>
Natural disaster victims aided
</TITLE>
<DESCRIPTION>
The description is usually a few sentences describing the cluster.
</DESCRIPTION>
<NARRATIVE>
The narrative often describes exactly what the user is looking for in the summary.
</NARRATIVE>
</QUERY>
```

Figure 4: Query object

```
<MEAD-CONFIG TARGET='GA3' LANG='ENG' CLUSTER-PATH='/clair4/mead/data/GA3'
  DATA-DIRECTORY='/clair4/mead/data/GA3/docsent'>

<FEATURE-SET BASE-DIRECTORY='/clair4/mead/data/GA3/feature/'>
  <FEATURE NAME='Centroid' SCRIPT='/clair4/mead/bin/feature-scripts/Centroid.pl HK-WORD-enidf ENG'/>
  <FEATURE NAME='Position' SCRIPT='/clair4/mead/bin/feature-scripts/Position.pl'/>
  <FEATURE NAME='Length' SCRIPT='/clair4/mead/bin/feature-scripts/Length.pl'/>
</FEATURE-SET>

<CLASSIFIER COMMAND-LINE='/clair4/mead/bin/default-classifier.pl \
  Centroid 1 Position 1 Length 9' SYSTEM='MEADORIG' RUN='10/09'/>

<RERANKER COMMAND-LINE='/clair4/mead/bin/default-reranker.pl MEAD-cosine 0.7'/>

<COMPRESSION BASIS='sentences' PERCENT='20'/>

</MEAD-CONFIG>
```

Figure 6: Mead-config object

12. References

Carbonell, Jaime G. and Jade Goldstein, 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Alistair Moffat and Justin Zobel (eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Melbourne, Australia.

Otterbacher, Jahna C., Dragomir R. Radev, and Airong Luo, 2002. Revisions that improve cohesion in multi-document summaries: a preliminary study. In *Proceedings of the Workshop on Automatic Summarization (including DUC 2002)*. Philadelphia: Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Radev, Dragomir, 2000. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In *Proceedings, 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong.

Radev, Dragomir, Sasha Blair-Goldensohn, and Zhu Zhang, 2001a. Experiments in single and multi-document summarization using MEAD. In *First Document Understanding Conference*. New Orleans, LA.

Radev, Dragomir, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Çelebi, Hong Qi, Elliott Drabek, and Danyu Liu, 2002. Evaluation of text summarization in a cross-lingual information retrieval framework. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.

Radev, Dragomir R., Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan, 2001b. Interactive, domain-independent identification and summarization of topically related news articles. In *5th European Conference on Research and Advanced Technology for Digital Libraries*. Darmstadt, Germany.

Radev, Dragomir R., Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan, 2001c. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Human Language Technology Conference (Demo Session)*. San Diego, CA.

Radev, Dragomir R., Weiguo Fan, and Zhu Zhang, 2001d. Webinessence: A personalized web-based multi-document summarization and recommendation system. In *NAACL Workshop on Automatic Summarization*. Pittsburgh, PA.

Radev, Dragomir R., Hongyan Jing, Malgorzata Stys, and Daniel Tam, 2004. Centroid-based summarization of multiple documents. *Information Processing and Management, in press*.

Radev, Dragomir R., Jahna Otterbacher, Hong Qi, and Daniel Tam, 2003. Mead reducs: Michigan at duc 2003. In *Proceedings of DUC 2003*. Edmonton, AB, Canada.

Witten, Ian H. and Eibe Frank, 2000. *Data Mining: Practical machine learning tools with Java implementations*. San Francisco: Morgan Kaufmann.

Zhang, Zhu, Sasha Blair-Goldensohn, and Dragomir Radev, 2002. Towards CST-enhanced summarization. In *Proceedings of the AAAI 2002 Conference*. Edmonton, Alberta.