Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Detecting Information-Dense Texts in Multiple News Domains

Yinfei Yang
Amazon Inc.
yinfyang@amazon.com

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Abstract

We introduce the task of identifying information-dense texts, which report important factual information in direct, succinct manner. We describe a procedure that allows us to label automatically a large training corpus of New York Times texts. We train a classifier based on lexical, discourse and unlexicalized syntactic features and test its performance on a set of manually annotated articles from business, U.S. international relations, sports and science domains. Our results indicate that the task is feasible and that both syntactic and lexical features are highly predictive for the distinction. We observe considerable variation of prediction accuracy across domains and find that domain-specific models are more accurate.

Introduction

Pioneers in natural language processing focused exclusively on the problem of semantic interpretation of text, developing methods for deriving formal representation of what a text is about. Their modern counterparts have instead zeroed in on issues of text style and potential impact, developing methods for deducing how information is conveyed and how information will be perceived by readers (Yu and Hatzivassiloglou 2003; Danescu-Niculescu-Mizil et al. 2009; Cook and Hirst 2013). Many of these, including assessing readability, helpfulness and trustworthiness of texts, have directly contributed to improvements in information access on the web (Kim et al. 2012; Agichtein et al. 2008; Pasternack and Roth 2011).

In our work we develop techniques for detecting information-dense news texts. Information-dense texts report important factual information in direct, succinct manner. The intuition behind our approach is to exploit conventions in journalistic writing in order to obtain samples of information-dense texts and texts that contain little important factual information. Specifically we base our analysis on the opening paragraph, called *lead*, of news articles. The purpose of the lead paragraph is to entice the reader to read the full article. News reports often adhere to the inverted pyramid structure, in which the lead conveys what happened, when and where, followed by more details that are less important. When writers adhere to this style of writing, the leads provide true examples of information-oriented style of writing, covering important factual information. Alternatively, however, the lead may be creative, provocative or entertaining rather than informative, as shown in the example below.

When the definitive history of the Iraq war is written, future historians will surely want to ask Saddam Hussein and George W. Bush each one big question. To Saddam, the question would be: What were you thinking? If you had no weapons of mass destruction, why did you keep acting as though you did? For Mr. Bush, the question would be: What were you thinking? If you bet your whole presidency on succeeding in Iraq, why did you let Donald Rumsfeld run the war with just enough troops to lose? Why didn't you establish security inside Iraq and along its borders? How could you ever have thought this would be easy?

The answer to these questions can be found in what was America's greatest intelligence failure in Iraq – and that was not about W.M.D.

Predictions of this dimension of writing style has immediate applications in summarization, question answering and information retrieval. Being able to predict if a lead is informative or creative will be particularly useful in automatic summarization of single news articles. It is well-established that in the news domain the lead of the article forms a great summary (Mani et al. 2002), which is indeed the intention when the writer follows the inverted pyramid style. If the lead is more creative, however, a fact based summary would be superior to the opening paragraph. The results from large-scale evaluations have confirmed that human summaries easily outperform the baseline of selecting the beginning of the article as a summary but automatic approaches to summarization cannot perform significantly better than this baseline on manual evaluation metrics (Nenkova 2005).

In this paper we introduce an automatic approach for labeling if a lead is information-dense or not and train a supervised classifier for the distinction. We evaluate our model on manually annotated data, showing that domain-specific models are much more accurate than general models of informativeness. Lexical and syntactic features appear to be particularly well suited for the task. Our analysis of manual annotations reveals that non-informative article leads are common and that our supervised approach comfortably outperforms a baseline of assuming that each newspaper lead is informative.

Data

We use summarization-inspired heuristics to acquire a large set of indirect labels, to use as training data for an informativeness classifier. The data for our experiments comes from the New York Times (NYT) corpus (LDC2008T19). This corpus contains 20 years worth of NYT, along with metadata about the newspaper section in which the article appeared and manual summaries for many of the articles. We selected a subcorpus of 30,614 articles published in 2005 or 2006 for our experiments.

We expect that the degree to which a text would be judged to be information-dense will be influenced by the genre of the article. To study cross-genre differences in the characteristics of informative texts, we select articles from four distinct genres: Business, Sports, Science and US International Relations (or Politics for short).

We use the article/summary pairs in the original NYT corpus to induce a labeling for informativeness on a subset of the data. Manual summaries convey the most important factual information in the article. High similarity between the lead and the summary therefore would indicate that the lead is informative while low similarities suggests that the lead is not informative.

For articles with manual summaries of at least 25 words, we calculate an informativeness score. The score is computed as the fraction of words in the lead that also appear in the abstract.

It is reasonable to expect that our indirect annotations will be noisy. To obtain cleaner data for training of our model, for each genre, we only use the leads with scores that fall below the 20th percentile and above the 80th percentile. In the general model, percentiles were determined on all data, without regard of domain. For domain-specific models, the distribution of abstract overlap scores reflected only the data from the particular domain.

Finally, our balanced corpus of informative and non-informative leads contains 2,256 from the Business section, 631 from Science, 588 from Sports and 1,119 from US international relations (politics).

Below we show some examples of leads that were labeled by this approach as informative and non-informative respectively, from the business and sports domains.

Informative:

[Business] *The European Union's chief trade negotiator, Peter Mandelson, urged the United States on Monday to reduce subsidies to its farmers and to address unsolved issues on the trade in services to avert a breakdown in global trade talks.*

Ahead of a meeting with President Bush on Tuesday, Mr. Mandelson said the latest round of trade talks, begun in Doha, Qatar, in 2001, are at a crucial stage. He warned of a "serious potential breakdown" if rapid progress is not made in the coming months.

[Sports] *Jack Snow, an outstanding receiver for the Los Angeles Rams for more than a decade and a longtime radio analyst for the Rams franchise, died Monday at a hospital in St. Louis. He was 62.*

The cause was complications of a staph infection, the Rams said. Snow had double hip replacement surgery last spring. The Rams' internal medicine physician, Douglas Pogue, said last week, according to The Associated Press, that the staph condition originated as a sinus infection, then entered the bloodstream and infected an artificial hip joint.

Non-informative:

[Business] *"ART consists of limitation," G. K. Chesterton said. "The most beautiful part of every picture is the frame." Well put, although the buyer of the latest multimillion-dollar Picasso may not agree.*

But there are pictures – whether sketches on paper or oils on canvas – that may look like nothing but scratch marks or listless piles of paint when you bring them home from the auction house or dealer. But with the addition of the perfect frame, these works of art may glow or gleam or rustle or whatever their makers intended them to do.

[Sports] *Shortly after the Indianapolis Colts were defeated by the San Diego Chargers yesterday, Nick Buoniconti's telephone rang.*

"It was my old teammate Dick Anderson," Buoniconti said later from his home in Miami. "He's meeting me tomorrow for a Champagne toast."

Features

We explore a variety of lexical features, as well as discourse and unlexicalized syntactic features which may distinguish informative from non-informative writing.

Lexical features

Some of our features are derived from semantic dictionaries which encode salient properties of the words in a domain independent manner. We also experiment with raw statistics, identifying domain dependent features using pointwise mutual information.

MRC Database The MRC Psycholinguistic Database (Wilson) is a machine usable dictionary containing 150,837 words, different subsets of which are annotated for 26 linguistic and psycholinguistic attributes. We select a subset of 4,923 words normed for age of acquisition, imagery, concreteness, familiarity and ambiguity. We use the list of all normed words in a bag-of-words representations of the leads. This representation is appealing because the feature space is determined independently of the training data and is thus more general than the alternative lexical representation that we explore.

The value of each feature is equal to the number of times it appeared in the lead divided by the number of words in the lead. We observe the words with higher familiarity scores scores, such as *mother, money, watch* are more characteristic for the non-informative texts, and appeared to be among the best indicators for the class of the lead.

MRC Concreteness In the MRC the degree to which a word is associated with imagery, concreteness, familiarity and ambiguity are given as a score in a range. To develop a more detailed representation, we split the range for each property into 230 sub-intervals. Each interval corresponds to a feature and the value of the feature is the fraction of words that fall in this interval. This representation has proven successful in other applications of content analysis (Klebanov

and Flor 2013). We report this representation only for the Concreteness norms, which appeared to distinguish best between the two classes[1].

We observed that informative lead contain words with lower familiarity and high concreteness.

LIWC and General Inquirer Databases LIWC (Pennebaker and Francis 2007) and General Inquirer (Philip J. Stone 1966) are two dictionaries in which words are grouped in certain semantic or functional categories, which we call tags for brevity. LIWC contains 4,496 words and General Inquirer contains 7,444 words. We compute the histogram of distribution of tags normalized by the total number of words for each lead. LIWC and General Inquirer tags are treated as separate representations.

Mutual Information The representations described so far capture genre independent information about word properties. We also introduce a genre-dependent representation, using mutual information to measure the association between particular words and the informative and non-informative writing styles in the training data. For each domain, we compute the mutual information between words and lead informativeness (Church and Hanks 1990) .

We select the top 500 words with highest associations with each of the writing styles, for a total of 1,000 features. The value of the feature is 1 if the word occurs in the lead and 0 otherwise.

Syntactic Features

Syntactic features reflect the way in which content is structured. Recent work has demonstrated for example that models of local coherence based on syntactic patters are capable of distinguishing if an article is published in a top tier or lower tier venue (Louis and Nenkova 2012). We experiment with the two types of syntactic representation compared in that work and expect that syntactic features may contain valuable cues as to whether a lead is information-dense or not.

Production Rules In this representation, we view each sentence as the set of grammatical productions, *LHS* $\rightarrow$ *RHS*, which appear in the parse of the sentence. We keep only non-terminal nodes in our work, excluding all lexical information, so the lexical and syntactic representations capture non-overlapping aspects of the writing style in a text.

d-sequence An alternative representation introduced by (Louis and Nenkova 2012) represents a sentence as a linear sequence of non-terminals which appear at the same depth d in the constituency parse of the sentence. In our work we choose d to be one greater than the depth of the main verb of the sentence. We computed 1- to 4-grams over of d-sequences. We evaluated classifiers trained on each order n-gram separately. The size of gram makes little difference to the performance, so for simplicity of presentation, we report only the results for unigram features.

[1]We did experiment with the other dimensions and results were only slightly lower.

Discourse Features

We compute features that capture the flow between sentences in the lead in terms of discourse relations and entity mentions. Prior work has shown that such features capture aspects of text readability and coherence (Barzilay and Lapata 2008; Lin, Ng, and Kan 2011), but no prior work has tested the extent to which such features correlate with writing style. We would expect that the more informative texts will have more clear entity structure and more explicit discourse relations.

PDTB Style Discourse Relations To label discourse relations, we employ the end-to-end PDTB-style discourse analysis tool of (Lin et al. 2012). It labels both explicit discourse relations that are signaled by a discourse connective and implicit relations that are inferred by the reader even when a connective is missing. Leads are represented in terms of the fraction of each type of discourse relation present in the text. The tool detects Contingency, Comparison, Temporal and Continuation relations, as well as more narrow subclasses of each of these relations. In our experiments, results using the fine-grained classes and those based only on the four main discourse relation classes differredminimally, so for simplicity we keep only these general relation types for the final evaluation results.

Entity Grid Entity grid is a 2D array representing grammatical roles for entities in the sentences in text. Each grid cell contains the grammatical role of the entity in the specific sentence, where the grammatical roles are subject (S), object (O), neither subject nor object (X) and absence from the sentence (-). Global transition patterns of grammatical roles in adjacent sentences, e.g. SO, XS, O-, reflect the entity coherence of the text.

We use the Brown Coherence Toolkit (V1.0) (Elsner and Charniak 2011) to generate the entity grid array. Some entities only occur few times because leads are usually short, so we only keep the columns coresponding to head nouns. Then we compute the distribution of transition patterns of head nouns to form 16 entity grid features for each lead.

Evaluation on automatically labeled data

We trained a binary classifier using LibSVM (R.-E. Fan and Lin 2008) with linear kernel and default parameter settings. We perform 10-fold cross-validation on the automatically labeled data with all features combined, but also analyze the performance when only a given class of features is used.

The results are presented in Table 1. Because of the way data was labeled, the two classes are of equal size, with 50% accuracy as the random baseline.

Each row in the table corresponds to a system trained with only the specified features. The final row shows the results for a classifier using all features and these uniformly lead to the best results.

The columns correspond to the domains we study. The domain-specific models were trained and tested only on the data from the given domain and the results are shown in the first four columns. The general model is trained and tested on the combined data and its performance is shown in the last column. Depending on the domain, accuracies are high, ranging between 85% for business and 73% for sports.

Interestingly, of the lexical features, the corpus-independent representation using the list of words from the MRC database (s_1) has the best performance for all domains. The corpus-dependent lexical representations based on mutual information between word and the text-informativeness class (s_5) have much lower performance, by almost 7% for the science domain for example. The fact that the representation designed independently of the training data can lead to such good results is a positive finding, indicating that the results are likely to be robust.

On par with the MRC lexical representations are the production rule syntactic representations, with accuracy over 80% for the business and science domains. In contrast the $d-$sequence syntactic representation does not appear to be that good for the task and the classifier that uses this representation performs markedly worse than that based on production rule representations of syntax.

Discourse features appear to be weak predictors, in some domains barely beating chance. This may be due to the fact that the leads are relatively short, so that discourse features are sparse. It is also quite possible that local coherence in information-dense text is similar to those that are not.

Uniformly across domains, MRC lexical representations and production rule syntactic representations are the best, while alternative lexical representation and discourse representations are comparatively weak. It will be of interest to know to what extent the MRC lexical and the production rule syntactic representations make mistakes on different test instances or whether they are in fact fully comparable and behave similarly.

We leave the issues of more sophisticated classifier combination and feature selection for future work. These are likely to improve both performance and our understanding of the characteristics of information-dense texts. Next, we turn to evaluation of the classifiers on manually annotated data, to complete our feasibility study for the task.

Evaluation on manual annotations

So far we have established that recognition of information-dense texts can be done very accurately when the label for the lead is determined on the basis of intuitive heuristics on the available article/summary resources. We would like however to test the models on manually annotated data as well, in order to verify that the predictions indeed conform to reader perception of the style of the article.

The authors of the paper manually annotated a set of 400 NYT articles, 100 from each domain, with human judgements of informativeness. Similar to prior work on grammatically judgements (E.G. Bard and Sorace 1996; Cowart 1997), the annotation was done with respect to a reference article that fell around the middle of the informativeness spectrum. The annotator gave both a categorical label for the article (less informative or more informative than the reference) and a real values score via a sliding bar. The categorical labels were used to test classification models, the real-valued ones were used to compute correlation with the classification score produced by the classifier. The annotated articles were randomly picked from the NYT data and did not appear in the data for which we reported cross-validation experiments in the previous section and which we use as training data for the classifiers that we evaluate here.

Articles were labeled by domain, all 100 articles from the same domain grouped together and displayed in random order. The reference article in each case was drawn from the respective domain.

	Business	**Science**	**Sports**	**Politics**	**General**
$\mathbf{s_1}$	0.820	0.796	0.699	0.765	0.830
$\mathbf{s_2}$	0.778	0.776	0.690	0.743	0.771
$\mathbf{s_3}$	0.753	0.691	0.650	0.711	0.763
$\mathbf{s_4}$	0.750	0.723	0.654	0.717	0.776
$\mathbf{s_5}$	0.796	0.728	0.651	0.750	0.814
$\mathbf{s_6}$	0.783	0.643	0.641	0.711	0.767
$\mathbf{s_7}$	0.828	0.803	0.702	0.757	0.822
$\mathbf{s_8}$	0.685	0.524	0.582	0.668	0.654
$\mathbf{s_9}$	0.631	0.618	0.563	0.625	0.646
f	**0.851**	**0.816**	**0.733**	**0.783**	**0.846**

Table 1: Binary classification accuracy of 10-fold cross validation on the automatically labeled set for different classes of features: MRC Dataset (s_1), MRC Concreteness (s_2), LIWC (s_3), General Inquirer (s_4), Mutual Information (s_5), d-sequence (s_6), Productions (s_7), PDTB Discourse (s_8), Entity Grid (s_9), and all features combined (f). Domain-specific models are trained and tested only on data from the same domain, the general model uses all domains combined.

Inter-annotator agreement All 400 test leads were annotated as being information-dense or not and with a real-value indicator of the extent to which they are information-dense.

Table 2 shows the percent agreement between the two annotators, as well as the correlation of the real-value annotation of informativeness. For the binary annotations we also compute the Kappa statistic.

The agreement is high for all domains. It is highest, almost 80%, on the political articles and lowest, 70%, on the business articles. The correlations of real-value degree of informativeness exceed 0.5 and are highly significant for all domains.

Kappa however is relatively low, indicating that the annotation task is rather difficult. To refine our instructions for annotation, we adjudicated all articles for which there was no initial agreement on the label. Both authors sat together, reading the reference article and each of the leads to be annotated, discussing the reasons why the article could be labeled information-dense or not. In many cases, the final decision was made by taking into account the sub-domain of the article, as well as the reference article for the specific genre.

	Agreement	Kappa	Correlation
Business	0.70	0.405	0.608
Science	0.74	0.455	0.523
Sports	0.73	0.460	0.522
Politics	0.78	0.550	0.711

Table 2: Inter-annotator agreement on manual annotations. Percent agreement is computed on the binary annotation, correlation is computed on the real-value degree of informativeness of the article. All correlations are highly significant, with $p < 0.001$.

Are Leads Informative? Table 3 shows the number of articles in each domain that were labeled as information-dense by each of the annotators, as well as in the combined set after adjudication. Because we annotated 100 articles in each domain, this is also the percentage of information-dense articles our of all annotated.

It is clear that the assumption prevailing in summarization research that the lead of the article is always information-dense is not reflected in the data we analyze here. The majority of articles in the politics domain, which are representative of the data on which large-scale evaluations of summarization systems tend to be performed, focused on specific current events, are indeed information-dense.

The second largest proportion of information-dense leads is in the business domain. There the articles are often triggered by current events but there is more analysis, humor and creativity. In these leads important information can often be inferred but is not directly stated in factual form.

In sports the factual information that the text needs to convey is not much and it is embellished and presented in a verbose and entertaining manner. In the science journalism section many leads only establish a general topic or an issue, or include a personal story about someone related to the scientific topic of the article. Particularly the second annotator considered less than a third of these articles to be information-dense.

	Combined	Anno_1	Anno_2
Business	53	47	57
Science	37	45	27
Sports	49	45	50
Politics	61	55	61

Table 3: Percentage/number of information-dense articles in the combined and individual manual annotations.

Classifier evaluation Here we evaluate the classifier trained on heuristically labeled data (with all features) on the manual annotations. Results from the domain-dependent and the overall general model are shown in Table 5. Accuracy computed against each of the two individual annotations is shown in the last two columns of the table. Prediction accuracies are very similar regardless of which annotator provides the labels. As in the heuristically labeled data, recognition accuracies are higher for the politics and business domains (around 75%) and lower for the science and sports domains (around 67%).

We also evaluate the prediction on the combined labels, as well as individually on the subsets of the data for which the two annotators agree on the label in the first stage of annotation and those for which adjudication was needed. Clearly, the classifier captures characteristics of information-dense information quite well. The accuracy on the subset of the data for which the annotators agree is much higher than that for individual annotators, indicating that when the text has mixed characteristics leading the annotators to disagree, it is more likely that the classifier makes more errors as well.

On the agreed test set, accuracy is around 85% for the politics and business domains, and in the lower seventies for sports and science. On the combined set, accuracies are much higher than the baseline of predicting that all leads are informative.

It is clear that the domain-specific models are much more accurate for the politics and business domains. The simpler general model works equally well for the science and sports domains. Further experiments are needed to elucidate the reasons for this result. The science and sports domains appear to be more difficult, with lower inter-annotator agreement and more leads that are not information-dense. At the same time these two domains also had less training data available, so results may improve if more data is labeled. Yet another reason may be that the labels in the manual dataset were domain-specific, with each article labeled as being information-dense or not with respect to a reference of the same domain, and with the knowledge of the same domain. In training however the labels for the overall domain-independent model were based on the degree of overlap between the human summary and the lead across all domains. In that case many of the articles from sports and science were labeled as being not information-dense. Future work will clarify what is the most advantageous approach for addressing the problem in order to develop a reliable classifier for use in practical applications.

Finally, we compute the correlation between the classification score from the SVM classifier and the real-value annotation of degree of informativeness by the two annotators. These are shown in Table 6. All correlations are highly statistically significant. In line with what we have seen in the analysis of other results and inter-annotator agreement, the correlation is highest for politics and business leads and lowest in the sports domain.

Similarly we compute the precision of prediction stratified according to the classifier confidence in that prediction. Figure 1. The precision of high confidence predictions is much higher than the overall accuracy, and confident predictions are made by the classifier for a large fraction of the test data. The trend suggests that automatic models for predicting text informativeness are likely to be more accurate if they predict real-value informativeness scores, as the classification scores of a classifier or in a regression model.

Conclusion

In this paper we introduced the task of detecting information-dense news article leads. We use arti-

Table 4: Classification accuracy on manual annotations

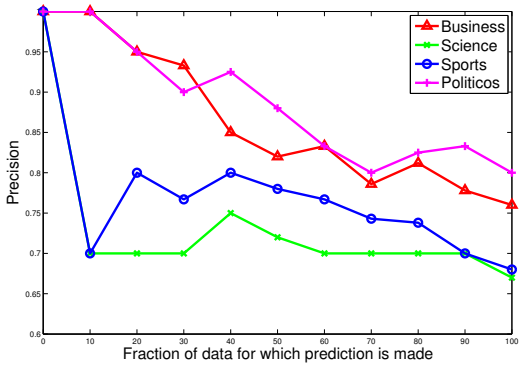
Business	Combined	Anno_1	Anno_2
Domain model	**.76** (**.843**; .567)	**.72**	**.75**
Overall model	.74 (.771; **.667**)	.64	.74
Baseline	.53 (.528; .533)	.47	.57
Science	Combined	Anno_1	Anno_2
Domain model	**.73** (**.730**; **.731**)	**.67**	**.67**
Overall model	.69 (.716; .615)	.65	**.67**
Baseline	.37 (.311; .538)	.45	.27
Sports	Combined	Anno_1	Anno_2
Domain model	**.70** (**739**; .593)	**.67**	**.68**
Overall model	**.70** (.685; **.741**)	.66	.61
Baseline	.49 (.465; .556)	.45	.50
Politics	Combined	Anno_1	Anno_2
Domain model	**.80** (**.859**; .591)	**.74**	**.74**
Overall model	.72 (.769; .545)	.70	.72
Baseline	.61 (.603; **.636**)	.55	.61

Table 5: Binary classification results on human annotated datasets for models trained on heuristically labeled data. In brackets we show accuracy on the agreed and adjudicated subsets of the test set respectively.

	Anno_1	Anno_2
Business	0.546	0.553
Science	0.471	0.498
Sports	0.357	0.389
Politics	0.638	0.554

Table 6: Correlation between predicted probabilities and human annotated scores. All correlations are highly significant with $p < 0.001$.

Figure 1: Predication precision based on probability ranking



cle/summary pairs from the NYT corpus to heuristically label a large set of articles as informative when the lead of the article overlaps highly with the human summary and as uninformative when the overlap is low.

We test lexical, syntactic and discourse features for the task. Syntactic production rule representations and corpus-independent lexical representations from a vocabulary defined by the MRC lexicon prove to be the most useful predictors of lead informativeness. The classifier that combines all features however works best. In the paper we presented detailed feature class evaluation in cross validation on heuristically labeled data and present results only for the model with all features on manual annotations for information-density. In results not reported in the paper we observe that evaluation on the manual annotations leads to identical conclusions.

Our analysis reveals that there is a large variation across news domains in the fraction of information-dense leads and in the prediction accuracy that can be achieved. Contrary to popular assumptions in new summarization, we find that a large fraction of leads are in fact not information-dense and thus do not provide a a satisfactory summary.

Overall, domain-specific models are more accurate than a general model trained on all data pooled together.

In this work, we have established the feasibility of the task of detecting information-dense texts. We have confirmed that the automatic annotation of data captures distinctions in informativeness as perceived by people. In future work the training set can be extended to include more of the NYT data. It is also of interest to characterize the features indicative of information-dense text, to develop better approaches for combining classifiers based on independent feature sets and to formalize the prediction in terms of real-value scores.

References

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, 183–194.

Barzilay, R., and Lapata, M. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1):1–34.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.* 16(1):22–29.

Cook, P., and Hirst, G. 2013. Automatically assessing whether a text is clichéd, with applications to literary analysis. *NAACL HLT 2013* 13:52.

Cowart, W. 1997. *Experimental Syntax: applying objective methods to sentence judgement*. Thousand Oaks, CA: Sage Publications.

Danescu-Niculescu-Mizil, C.; Kossinets, G.; Kleinberg, J.; and Lee, L. 2009. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, 141–150.

E.G. Bard, D. R., and Sorace, A. 1996. Magnitude estimation for linguistic acceptability. *Language* 72(1):32–68.

Elsner, M., and Charniak, E. 2011. Disentangling chat with local coherence models. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *ACL*, 1179–1189. The Association for Computer Linguistics.

Kim, J. Y.; Collins-Thompson, K.; Bennett, P. N.; and Dumais, S. T. 2012. Characterizing web content, user inter-

ests, and search behavior by reading level and topic. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, 213–222.

Klebanov, B. B., and Flor, M. 2013. Word association profiles and their use for automated scoring of essays. In *Proceedings of ACL*.

Lin, Z.; Liu, C.; Ng, H. T.; and Kan, M.-Y. 2012. Combining coherence models and machine translation evaluation metrics for summarization evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, 1006–1014. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, Z.; Ng, H. T.; and Kan, M.-Y. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 997–1006.

Louis, A., and Nenkova, A. 2012. A coherence model based on syntactic patterns. In *Proceedings of 2012 Joint Conference on Empirical Methods in Natural Language Processing*, 1157–1168. Association for Computational Linguistics.

Mani, I.; Klein, G.; House, D.; Hirschman, L.; Firmin, T.; and Sundheim, B. 2002. Summac: a text summarization evaluation. *Natural Language Engineering* 8(01):43–68.

Nenkova, A. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *AAAI*, 1436–1441.

Pasternack, J., and Roth, D. 2011. Making better informed trust decisions with generalized fact-finding. In *IJCAI*.

Pennebaker, J. W., B.-R. J., and Francis, M. E. 2007. Linguistic inquiry and word count: Liwc.

Philip J. Stone, Dexter C. Dunphy, M. S. S. D. M. O. 1966. The general inquirer: A computer approach to content anaysis.

R.-E. Fan, K.-W. Chang, C.-J. H. X.-R. W., and Lin, C.-J. 2008. Liblinear: A library for large linear classification. 9:1871–1874.

Wilson, M. The mrc psycholinguistic database: Machine readable dictionary. *Behavioural Research Methods, Instruments and Computer, version 2*.

Yu, H., and Hatzivassiloglou, V. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, 129–136.