Journal of Artificial Intelligence Research 60 (2017) 179-219 Submitted 12/16; published 09/17

Combining Lexical and Syntactic Features for Detecting Content-Dense Texts in News

Yinfei Yang YANGYIN7@GMAIL.COM
1600 Amphitheatre Pkwy
Mountain View, CA 94043

Ani Nenkova NENKOVA@SEAS.UPENN.EDU
University of Pennsylvania
3330 Walnut Street
Philadelphia, PA, 19103 USA

Abstract

Content-dense news report important factual information about an event in direct, succinct manner. Information seeking applications such as information extraction, question answering and summarization normally assume all text they deal with is content-dense. Here we empirically test this assumption on news articles from the business, U.S. international relations, sports and science journalism domains. Our findings clearly indicate that about half of the news texts in our study are in fact not content-dense and motivate the development of a supervised content-density detector. We heuristically label a large training corpus for the task and train a two-layer classifying model based on lexical and unlexicalized syntactic features. On manually annotated data, we compare the performance of domain-specific classifiers, trained on data only from a given news domain and a general classifier in which data from all four domains is pooled together. Our annotation and prediction experiments demonstrate that the concept of content density varies depending on the domain and that naive annotators provide judgement biased toward the stereotypical domain label. Domain-specific classifiers are more accurate for domains in which content-dense texts are typically fewer. Domain independent classifiers reproduce better naive crowdsourced judgements. Classification prediction is high across all conditions, around 80%.

1. Introduction

News articles are written with different goals in mind. Some aim to inform the reader about an important event, focusing on specific details such as who did what to whom where and when. Others aim to provide background information, facts related to an event and necessary to understand an event but not newsworthy by themselves. Yet others seek to entertain the reader, or to showcase the brilliant mastery of language and the wit of the author.

In this paper we introduce the task of detecting if a text is content-dense or not. Content-dense news report important factual information about an event, in direct and succinct manner. Prototypical examples of content-dense texts are newswire articles, which are usually perfect answers to a “What happened?” question, grounded in a specific event. In general news, however, newswire-like, content-dense text is not the norm.

We base our analysis on the opening paragraph, called the *lead or lede*, of news articles drawn from the New York Times. News reports often adhere to the inverted pyramid structure, in which the lead conveys what happened, when and where, followed by more details in the body. Information that is not essential is included in the final tail. When writers adhere to this style of writing, the leads are informative and provide positive examples of content-dense texts. Alternatively, the lead may be creative, provocative or entertaining rather than informative, providing examples of non content-dense texts.

Consider the leads below, from the politics and sports section of the New York Times. The first two are content-dense leads. The other two are non content-dense leads that do not focus on events; and which are much richer stylistically.

Content-dense:

*[**Politics**] Evo Morales, a candidate for president who has pledged to reverse a campaign financed by the United States to wipe out coca growing, scored a decisive victory in general elections in Bolivia on Sunday.*

Mr. Morales, 46, an Aymara Indian and former coca farmer who also promises to roll back American-prescribed economic changes, had garnered up to 51 percent of the vote, according to televised quick-count polls, which tally a sample of votes at polling places and are considered highly accurate.

*[**Sports**] North Carolina (29-1) and Duke (26-3) of the Atlantic Coast Conference received No. 1 seedings yesterday in the 64-team women's N.C.A.A. tournament, along with Ohio State (28-2) and Louisiana State (27-3).*

The top-ranked Tar Heels received the No. 1 overall seeding, but were placed in what appears to be the most difficult regional.

Non content-dense:

*[**Politics**] 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.

*[**Sports**] With his silver pants and dark blue jersey covered by a mottled mix of grass stains, paint and mud, New England Patriots running back Corey Dillon sat on an aluminum bench on the sideline at Gillette Field on Sunday, looking exhausted and frozen.*

Only a few minutes remained in the Patriots' 20-3 victory over the Indianapolis Colts, and Dillon was resting. He stared at the field, snowflakes swirling around his head as the realization of his first playoff victory swirled inside it.

Below we propose an approach for labeling short news texts as content-dense or not. Our analysis of manual annotations reveals that uninformative article leads are common. We investigate several types of lexical and non-lexicalized syntactic features for distinguishing content-dense texts from other more general or creatively written texts. We present a two-layer classifier model which significantly outperforms a baseline assuming that all news leads are content-dense. We also study the robustness of the definition of content density across domains, as well as the performance of domain-dependent and domain-independent (general) classifiers.

2. Motivation

Traditionally, natural language processing practitioners work under the assumption that the direct goal of text analysis is to ultimately derive a semantic interpretation of text. Our work deviates from this tradition and instead focuses on detecting style differences first, deferring or entirely foregoing semantic interpretation. This "style, then semantics if need be" approach to understanding reflects typical human behavior (Kahneman, 2011).

Under style we hope to capture *how* content is conveyed rather than exactly *what facts* are being communicated (Queneau, 1947) or what truth values one ought to assign to the expressed statements. This definition is remarkably close to decades-old attempts to define style as part of text typology:

> **Style** is used here to mean the way texts are internally differentiated other than by topic; mainly by the choice of the presence or absence of some of a large range of structural and lexical features.
>
> (Sinclair & Ball, 1996)

In the article we also investigate how broad topics (our news domains) interact with style, both in the way domain information influences people's style judgements and in the change of the structural and lexical indicators of style across domains.

In spirit, our work belongs to a growing body of research concerned with developing methods for deducing how information in longer text[1] is conveyed and how information will be perceived by readers (Yu & Hatzivassiloglou, 2003; Danescu-Niculescu-Mizil, Kossinets, Kleinberg, & Lee, 2009; Jurafsky, Ranganath, & McFarland, 2009; Ashok, Feng, & Choi, 2013; Cook & Hirst, 2013; Louis & Nenkova, 2014). Our effort is complementary, and cannot be compared directly, to work concerned with propositional meaning directly, such as event detection (Peng, Song, & Roth, 2016; Feng, Huang, Tang, Ji, Qin, & Liu, 2016; Nguyen & Grishman, 2016), veridicality (Saurí & Pustejovsky, 2009; de Marneffe, Manning, & Potts, 2012) or fact-checking (Vlachos & Riedel, 2014).

3. Corpus

The data for our experiments comes from the New York Times (NYT) annotated corpus (LDC Catalog No. LDC2008T19). The corpus contains 20 years worthy of NYT editions, along with rich meta-data about the newspaper section in which the article appeared and summaries produced by information scientists for many of the articles. The leads of articles are explicitly marked in the corpus, so extracting the relevant text for further analysis is straightforward.

[1] rather than individual words and sentences

In our previous proof-of-concept work (Yang & Nenkova, 2014), we selected a subcorpus of articles published in 2005 or 2006 from four different genres (business, U.S. international relations, science and sports). Given the selection criteria, the data in that prior work contained considerably fewer articles from the science and the sports domains compared to the other two domains. Moreover, the performance of the content-dense classifiers in the science and sports domains was notably worse than the other two domains, which could be explained either by the fact that these classifiers were trained on smaller datasets or by the intrinsic difficult of predicting content density in these two domains. To definitively resolve this question, and to benefit from the largest training dataset possible, we extend the corpus to the full NYT corpus in the experiments reported in this manuscript.

We also expect that the degree to which a text would be judged to be content-dense, reporting on important event in a direct manner, is influenced by the domain of the article. It is reasonable to expect that typical events in science or sports would not be considered of the same importance as international political or business events. To study the cross-domain differences, we analyze four news domains: Business, Sports, Science[2] and US International Relations (or Politics for short).

3.1 Training Set Heuristic

To automatically label leads as content-dense or not, we make use of the manual summaries which accompany many articles in the NYT corpus. For the articles with content-dense leads, the manual summary will be very similar to the lead itself, as this type of lead by definition provides a fact-focused summary of the article. For leads that simply seek to engage the reader via more creative devices, the manual summary will differ considerably from the lead. Overall, the similarity between the lead and the manual summary provides a strong indication of the importance and factual, event-oriented, nature of the information expressed in the lead.

For articles with manual summaries of at least 25 words, we calculate a content-dense score. For each word in the summary, a tuple $t(w, pos)$ is created containing the word and its part of speech. The score is computed as:

$$Score = \frac{\text{\# of t(w, pos) also in leads}}{\text{\# of t(w, pos) in sum}} \tag{1}$$

3.2 Label Analysis

Table 1 shows details about the number of all NYT articles from each of the four domains. The first column shows the number of articles in the NYT from the given domain. The second column shows the number of articles used for training domain-dependent classifiers (we explain the selection procedure below). Overall only about one third of articles have associated manual summaries.

The distribution of content-dense scores assigned as a function of the overlap with the human summaries is shown in Figure 1. In the business domain the distribution of scores

[2] The science articles are from the CATS corpus (Louis & Nenkova, 2013), which only contains articles published after 1999.

Table 1: Number of articles in the corpus.

	Total number of articles	*Articles used in training (Percentage)*
Business	149,113	21,224 (14.2%)
Science	23,240	7,737 (33.%)
Sports	134,925	10,670 (7.9%)
Politics	45,926	10,503 (22.8%)
Overall	353,204	50,134 (14.2%)

is almost uniform, reflecting the fact that in that section there are articles about important events—company mergers, unexpected stock price changes, product announcements and lawsuits—but also non-event specific analysis of current trends, minor events such as auctions and people-centered pieces about prominent business men and women.

In sports and science, the distribution of content-dense scores is clearly skewed towards the non content-dense end of the spectrum. In these domains writers more often resort to the use of creative and indirect language meant to provoke readers' interest.

The content-dense scores in politics is almost normally distributed, with mean roughly in the middle of the possible range, and much higher than any of the other domains. The non content-dense leads in this domain usually provide a commentary on an ongoing event rather than reports of a specific new development.

In the rest of the paper, we focus on the binary classification task of predicting if a lead is content-dense or not. However, it is reasonable to expect that our indirect labeling scores are noisy. To obtain cleaner data for training our model, we label only the leads with most extreme scores: we assign the label non content-dense to the leads with scores that fall below the 20th percentile and label content-dense to leads that score above the 80th percentile for their domain. The 20th/80th percentile sets are colored red in Figure 1. In the general (domain-independent) model, the data is pooled together and again the leads with lowest scores are assigned to the non content-dense class and the leads with highest scores are considered content-dense.

4. Methodology

In this section, we introduce the features and models we used in our experiments. In our prior experiments (Yang & Nenkova, 2014), we found that lexical features are well-suited for the task, particularly lexical representations determined independently of the training data. Along with these, unlexicalized syntactic representations also lead to remarkably good results. A number of other representations we experimented with did not appear to be that beneficial for the task. Motivated by these findings, here we study in depth the lexical representations and the unlexicalized syntactic representation, and explore ways to combine the predictions of these models to achieve even better accuracy.

4.1 Features

We compare and combine two lexical and one syntactic representation. For the lexical representation, we use the vocabulary from the **MRC Database(MRC)**, which is inde-

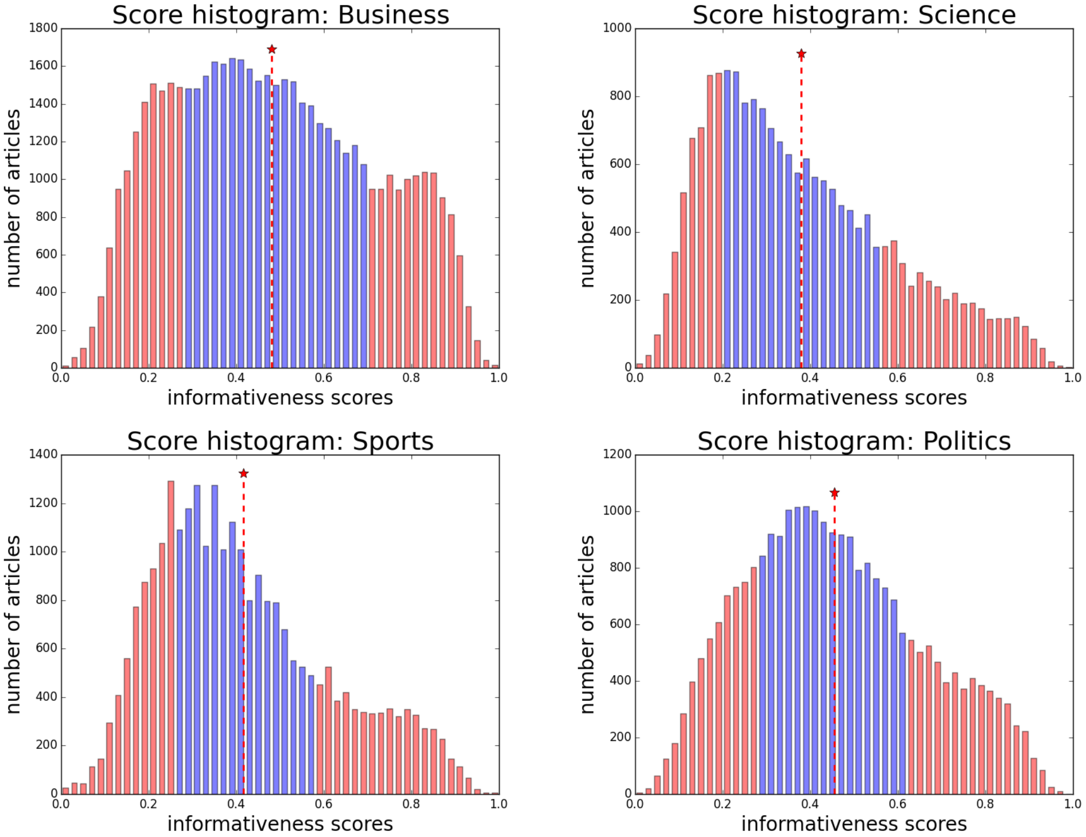


Figure 1: Score histograms for the four genres: [**Top Left**] Business, [**Top Right**] Science, [**Bottom Left**] Sports, [**Bottom Right**] Politics. 20th and 80th percentiles are colored red. Red star indicates the average content-dense score for each genre.

pendent of our training set and a vocabulary derived from the training set and weighted by **Mutual Information(MI)**. The syntactic representation is simply the list of **Production Rules(PR)** from the constituency parse of the sentences in the lead.

4.1.1 MRC Database(MRC)

The MRC Psycholinguistic Database (Wilson, 1988) is an electronic 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. In (Wilson, 1988), the words were chosen among those with medium frequency in a large corpus and experiment subjects were asked to rate on a scale the degree to which each word has one of these properties. The MRC dictionary is a compilation of results from different studies, run by different research groups, with different criteria for selecting the list of words for which to solicit norms. We use the list of words which have at least one of above ratings. 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.

About 90% of the MRC vocabulary (4,647 words) appears at least once in the training data. About 4,300 appear more than five times.[3]

4.1.2 Mutual Information(MI)

The lexical representation described above is domain independent, determined without any knowledge about the data which will be used for training and testing of our classification models. We also introduce a domain-dependent lexical representation, derived from the training data for the classifiers and using mutual information to measure the association between particular words and the content-dense and non content-dense writing styles. For each genre, we compute the mutual information between words and lead type in the training data as:

$$\mathrm{MI_c} = \log \frac{p(\mathrm{word, c})}{p(\mathrm{word})p(\mathrm{c})} \tag{2}$$

Here c is either the content-dense or the non-content dense class. We only compute the MI scores for words that appear at least 5 times in the training set. 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.

The words with highest mutual information[4] with the content-dense classes and non content-dense classes are listed in Table 2.

The words with high mutual information with the content-dense class are distinctly domain specific. Content-dense leads in the business domain are more likely to talk about companies and their executives, deals, agreements and offers. Content-dense leads in science are more likely to discuss a specific study or drug, since they are overwhelming biased towards health-related topics. Sports content-dense leads are associated with specific sport events or deals. In politics, content-dense leads discuss American involvement and attacks.

In addition, the words “yesterday” and “today” also appear among those associated with content-dense leads, providing a strong indicator that the news is focused on a specific recent event rather than a general discussion or personal aspects story. The words associated with the non content-dense class in contrast tend to be related to non-specific activities (*find, feel, hear, smile, remember, sit, wait*) and focused on personal aspects () rather than on the professional roles (*man, people, friend, husband, guy, kid, child, friend*).

Only about half of the words in the mutual information representation also appear in the MRC.

[3] As we mentioned in the opening of this section, in our early work (Yang & Nenkova, 2014) we also experimented with other dictionaries, including LIWC and the General Inquirer. Results consistently confirmed that the MRC lead to best prediction results. This is also the largest resource, guaranteeing the best coverage of features for new texts. For these reasons, we include only MRC features in the work presented here, to focus the presentation on the cross-domain differences and classifier combination, which are novel with respect to our prior work.

[4] We ran 10 fold cross validation in the experiments. The mutual information is computed separately based on the training set of each fold. The words listed in Table 2 are from fold 0. High mutual information words from other folds are very similar.

Table 2: Top 30 selected words for each domain and overall data

	Content-dense	*Non content-dense*
Business	*company, yesterday, million, billion, today, percent, group, announce, executive, plan, share, corporation, york, part, deal, agree, largest, unit, court, agency, inc., commission, bank, include, firm, chief, agreement, chairman, offer, service*	*day, stock, ago, work, thing, good, investor, year, find, turn, long, man, economy, job, people, home, street, room, time, rate, lot, index, city, sit, mr., market, wall, money, ms., life*
Science	*study, health, today, yesterday, report, drug, official, research, federal, state, scientist, administration, disease, researcher, company, government, accord, human, virus, university, group, million, expert, announce, cell, include, cancer, united, agency, issue*	*day, mr., ms., ago, feel, hear, room, sit, walk, home, eye, life, friend, thing, run, talk, live, game, stand, back, family, hand, foot, good, morning, husband, hour, night, town, son*
Sports	*yesterday, today, league, team, national, million, season, association, official, die, cup, contract, race, year, deal, tonight, game, conference, major, president, round, lead, charge, announce, committee, victory, win, woman, world, series*	*fan, ago, watch, back, stand, ball, day, question, good, turn, moment, room, smile, feel, hand, time, wear, people, knicks, hear, remember, n.b.a., net, guy, sit, thing, stadium, shot, kid, walk*
Politics	*official, united, today, american, states, administration, mr., clinton, military, government, weapon, international, effort, attack, security, nuclear, force, report, intelligence, group, court, defense, nations, program, include, china, agency, secretary, nato, plan*	*man, day, world, war, people, time, u.s., ago, back, sit, thing, front, live, city, child, street, room, stand, saddam, morning, america, word, year, wait, car, kerry, young, friend, watch, hour*
Overall	*yesterday, today, company, million, official, billion, group, united, percent, announce, states, plan, administration, york, include, american, agency, government, federal, report, accord, court, executive, national, drug, part, state, international, corporation, deal*	*day, ago, thing, man, good, stock, time, room, sit, back, stand, turn, watch, street, hear, home, feel, people, long, life, lot, ms., walk, town, wall, word, friend, live, moment, eye*

4.1.3 Production Rules(PR)

Finally, we use production rules as the syntactic representation (Louis & Nenkova, 2012; Ganjigunte Ashok, Feng, & Choi, 2013; Post & Bergsma, 2013; Malmasi & Dras, 2014).

We view each sentence as the set of grammatical productions, $LHS \rightarrow RHS$, which appear in the syntactic parse tree of the sentence. We keep only non-terminal nodes, excluding all lexical information, so the lexical and syntactic representations capture non-overlapping aspects of writing style. All production rules from the training set are used in the representation. The numbers of production rules vary for the four domains, from 16,000 rules (Science) to 32,000 rules (Business).[5]

4.2 Classifier Combination

The three feature representations we introduced capture domain independent lexical clues for content-density, domain-dependent indicators for important events and general style of writing captured by the structure of sentences in the text. We train a logistic regression classifier with each class of features individually. Furthermore in this section, we examine two approaches for combining the predictions from the three classes of features.

4.2.1 Feature-Level Combination(C1)

First we examine the performance of feature-level combination to develop a system that makes use of all three types of indicators of content density. We concatenate the three feature representations together in a feature vector. The number of entries in the feature vector is equal to the sum of the number of features of the MRC, mutual information and production rule representations. Then we train a logistic regression model based on the concatenated feature representation. This way of combining evidence lead to overall improvements in our early work. However much work on ensemble learning has demonstrated that for variety of tasks this method of combination is not as powerful as decision-level combination (for example see Raaijmakers, Truong, & Wilson, 2008; van Halteren, Zavrel, & Daelemans, 1998; Metallinou, Lee, & Narayanan, 2010; Bertolami & Bunke, 2006). We treat the feature-level combination as the baseline for our experiments. Figure 2 (a) shows the structure of feature-level combination classifier.

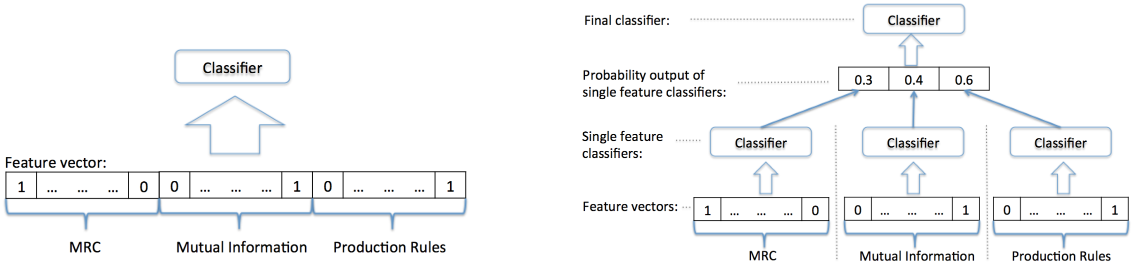
4.2.2 Decision-Level Combination(C2)

Classifier combination has been shown to outperform feature combination in a single classifier (Tulyakov, Jaeger, Govindaraju, & Doermann, 2008). There are multiple reasons why this may be the case, especially for a linear classifier like the one we use. Concatenating all features in a single representation makes the system prone to over-fitting, as the number of features becomes closer to the number of training examples. If the number of features of a given type is considerably smaller (for example there are many more features in the production rule representation compared to the mutual information representation), the signal contributing to the final decision may be dominated by the larger class, defeating the purpose of evidence combination. It could also lead to the presence of correlated features, for example in the combination of the two types of lexical features.

We propose a two layer classifier combination system. We first train a logistic regression classifier with each of the three feature representations individually. Then another model is trained, in which the features are the probabilities of the content-dense class from the first layer classifiers. In the experiment, the corpus is split into training set, development set and testing set. The first layer classifiers are trained on the training set, and the second

[5] Stanford CoreNLP package (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014) is used to extract production rules.

layer classifier is trained on development set. Figure 2 (b) illustrates the structure of the decision-level combination system.



(a) Feature-level combination

(b) Decision-level combination

Figure 2: Illustration of feature-level combination and decision-level combination

5. Evaluation on Automatic Annotations

In this section we will evaluate the effectiveness of each of the features as well as the two combination systems.

5.1 Classifier Evaluation

In the feature-level combination system, we train the binary classifier using Liblinear (R.-E. Fan & Lin, 2008) with L2-regularized logistic regression model setting. In the decision-level combination experiments, we first train binary classifiers based on each feature representation using LibLinear with the same settings. Using the probability outputs (for the content-dense class) of the first stage classifiers as features, we then train a final binary classifier using LibSVM (Chang & Lin, 2011) with linear kernel. Grid search is used on training and development set to find the best hyper-parameters in all models.

We perform 10-fold cross-validation experiments on the entire heuristically labeled data. The entire dataset is split into 10 partitions. At each run, five partitions are used for training first-stage classifiers and the feature-level combination classifier. Four partitions are used for training the second-stage combination classifier, which uses only the probabilities of the content-dense class from the first stage classifiers. One partition is used for testing the classifiers. We evaluate the two combination models on the automatically labeled data but also analyze the performance when only a single class of feature is used. [6]

The results are presented in Table 3. Because of the way the data was labeled, the two classes are of equal size, with 50% accuracy as the random baseline. The top three rows in the table corresponds to a system trained with only one class of features. The last two rows shows the results for the two combination systems. The columns correspond to the domains we study—business, science, sports and politics. The domain-specific models were

[6] We train the first- and second-stage classifiers on different portions of the training data in the fold in order to obtain realistic predictions from the first-stage classifier. If we were to use all nine partitions in the current fold, the second-stage classifier would be trained on the predictions of the first-stage classifier *on its own training data* which would be unrealistically accurate. The training protocol we adopt reflects better realistic usage of the combined classifier.

trained and tested only on the data from the given domain and the results are shown in the first four columns. The general, domain-independent model is trained and tested on the combined dataset and the last column shows its performance.

Precision, recall, F-score and accuracies are shown in the table. Depending on the domain, accuracies are high, ranging between 87.2% for business and 83.6% for politics. The precision and recall are very balanced according the numbers, which leads F-score very close to accuracy in all experiments. Here we mostly focus our discussion on accuracy.

Of the individual feature classes, the production rules representation leads to the best overall accuracy. Combining the representations at the feature level leads to improvement over the production rule classifier for the business and politics domain, as well as in the general domain-independent classifier but not for science and sports where performance using all features is in fact worse than using production rules alone.

In line with our previous work, all single feature classifiers have very good performances. The production rules (PR) syntactic representations lead to the best performance for all domains, with accuracies over or close to 80% for all domains. The most important rules are quite different in each genre, but the discovered patterns are mostly aligned with our intuition. For example, *VP ->VB NP PRT ADVP* is often associated with content-dense leads in Business, the example text like *VP ->VB[push] NP[the Czech currency] PRT[up] ADVP[sharply]*. The rule *NP ->JJ CD NNS*, however, is usually associated with non content-dense leads, e.g. *NP ->JJ[pre-April] CD[15] NNS[blues]*. The production rules with highest weights are listed in Appendix B.

Of the lexical representations, the MRC representation leads to better results, with accuracies varying from 82.7% for the business domain to 79.4% for the sports domain. The corpus-dependent lexical representations based on mutual information has a slightly lower performance: the accuracies range between 81.9% for business and 78.1% for politics.

The results for the general classifier—which is trained and tested on data from the four domains pooled together—are similar. For this classifier leads may change their labels, for example a sports article whose content-dense score is in the 80th percentile of scores for sports may fall below the 20th percentile when all data is combined.

The fact that the representations 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.

For all the domains and general domain-independent data, decision-level combination considerably improves the performance compared to classifiers trained with only one of the representations. It is the most accurate among the five classifiers that we compare, with up to 3.8% performance gain in politics compared to the best single feature classifier.

The baseline combination system, feature-level combination, performs worse than the decision-level combination. One of the possible reasons is that given the increased number of features, this model may require more training data to reach its performance potential. We study this aspect of model development in section 5.3.

5.2 Combining Classifiers with Different Representations

Here we evaluate different possible combinations of feature types. We compare these possibilities for decision-level combination, which we already established works better than feature-level combination.

Table 3: Binary classification results of 10-fold cross validation on the automatically labeled set for different classes of features and two fusion models for all domains: [P]recision / [R]ecall / [F]score / [A]ccuracy (%)

	Business				**Science**			
	P	R	F	A	P	R	F	A
MRC	82.0	84.4	83.1	82.7	79.4	84.1	81.6	81.1
MI	80.0	85.1	82.5	81.9	76.9	85.0	80.7	79.8
PR	83.8	83.9	83.8	83.8	83.2	84.7	83.9	83.8
C1	85.8	85.3	85.5	85.5	80.4	83.7	82.0	81.4
C2	**87.9**	**86.4**	**87.1**	**87.2**	**87.5**	**87.0**	**87.2**	**87.3**
	Sports				**Politics**			
	P	R	F	A	P	R	F	A
MRC	78.6	80.9	79.7	79.4	76.8	82.3	79.4	78.5
MI	76.0	84.4	80.0	78.8	74.7	84.8	79.4	78.1
PR	82.1	83.3	82.7	82.6	79.4	80.4	79.9	79.8
C1	80.8	83.0	81.9	81.5	77.0	83.6	80.1	80.8
C2	**86.0**	**85.0**	**85.5**	**85.6**	**83.1**	**84.2**	**83.6**	**83.6**

Table 4: Binary classification results of 10-fold cross validation on the automatically labeled set for different classes of features and two fusion models for general: [P]recision / [R]ecall / [F]score / [A]ccuracy (%)

	General			
	P	R	F	A
MRC	81.2	83.4	82.3	82.0
MI	79.4	82.2	80.8	80.6
PR	83.5	83.5	83.5	83.5
C1	84.4	85.8	85.1	85.0
C2	**86.8**	**86.4**	**86.6**	**86.7**

The motivation to examine combinations of features is that not all features are available in all applications. Moreover concerns about run time may make syntactic features undesirable in certain settings, where syntactic parsing may not be feasible. Mutual information representations also require larger training data for each domain of interest, to compute the mutual weights for each feature. So we examine the effectiveness of combining different feature classes. The multilayer structure makes the decision-level fusion easier to add or remove features. Developers can simply train a classifier based on new features, then add them to the second layer without affecting existing single feature classifiers.

We show the results from evaluating three different classifier combinations: MRC+MI (lexical features only), MRC+PR (domain independent features only) and MRC+MI+PR (all features together).

The results are shown in Table 5. The top row in the table corresponds to the baseline, feature-level combination model with all three classes of features. Rows 2-4 correspond to

Table 5: Binary classification results of 10-fold cross validation on the automatically labeled set for different combinations of features for all domains: [P]recision / [R]ecall / [F]1 / [A]ccuracy (%)

	Business				**Science**			
	P	R	F	A	P	R	F	A
C1	85.8	85.3	85.5	85.5	80.4	83.7	82.0	81.4
MRC+MI	84.8	84.8	84.8	84.8	83.7	82.7	83.1	83.2
MRC+PR	87.2	86.1	86.6	86.7	86.5	86.6	86.6	86.6
MRC+MI+PR	**87.9**	**86.4**	**87.1**	**87.2**	**87.5**	**87.0**	**87.2**	**87.3**
	Sports				**Politics**			
	P	R	F	A	P	R	F	A
C1	80.8	83.0	81.9	81.5	77.0	83.6	80.1	80.8
MRC+MI	82.2	81.8	82.0	82.0	79.5	82.3	80.9	80.6
MRC+PR	84.6	84.7	84.6	84.7	82.2	83.5	82.8	82.7
MRC+MI+PR	**86.0**	**85.0**	**85.5**	**85.6**	**83.1**	**84.2**	**83.6**	**83.6**

Table 6: Binary classification results of 10-fold cross validation on the automatically labeled set for different combinations of features for general: [P]recision / [R]ecall / [F]1 / [A]ccuracy (%)

	General			
	P	R	F	A
C1	84.4	85.8	85.1	85.0
MRC+MI	83.7	83.3	83.4	83.5
MRC+PR	86.5	85.4	85.9	86.1
MRC+MI+PR	**86.8**	**86.4**	**86.6**	**86.7**

decision-level models with the three different classifier combinations. As in previous tables, the first four columns correspond to domain-specific models, and the last column shows the results for the general, domain-independent model. Combination classifiers based on all three features in decision-level combination still has the highest accuracy, showing that each of the three representations contributes to the improved performance of the classifier. The domain independent features, MRC+PR with decision-level combination shows a competitive results too, suggesting that the mutual information representation is the one that could be removed with least degradation in performance. The accuracies are just slightly lower than the best, 0.5% lower for the business domain for example.

The decision-level combination of lexical representations has lower performance then the other two decision-level combination models. The accuracies range between 84.5% for the business domain and 80.3% for the politics domain. The combination of the two lexical representation leads to better performance than using either of the individual features classes, suggesting that MRC+MI combination at the decision-level is a good alternative when syntactic features are not available.

5.3 Is the Training Data Enough?

We now discuss the impact of the training set size on classifier performance. We evaluate the relationship between classifier accuracy and the increasing of the number of training instances for each domain. We start with a training set of 100 articles, growing to 6,500 instances in the training data, increasing the training set with 100 randomly selected articles in each step. Accuracy is computed on the same testing set for each domain. As in our previous experiments, 10-fold cross validation is performed. For each fold, there is a dedicated test set, which means all cross-validation iterations used the same test set. The reported results are an average of the accuracies on the fixed test set in each fold.

Figure 3 shows the accuracy/size curve for each domain. Among the four genres, decision-level combination of all three features has the highest accuracy. The accuracy increases rapidly with the increase of training data when the number of training articles is less than 2,000. When the size is larger than 2,000, it continues to increase, but very slowly. The decision-level combination of MRC+PR features, which is the second best model for all domains, behaves similarly. The accuracy of the MRC+MI decision-level combination is the worst of the combination systems and exhibits the slowest increase.

The accuracies of decision-level combination with 6,500 training article are already very close to the final numbers with full training set (shown in table 5). Increasing the number of training instances barely changes the performance after this point.

The baseline, feature-level combination, has the lowest accuracies. Yet we still see increase in accuracy as the training set size increases. For three of the domains, its performance becomes the same as that of the MRC+MI combination with a large enough training set.

The results also indicate that decision-level combination is able to achieve better performance with less training data.

The graphs suggest that the difference in performance of the content-density predictor in the four domains likely reflects the difficulty of the domain rather than the difference in training data size.

6. Evaluation on Human Annotations

So far we have established that recognition of content-dense texts can be done very accurately when the label for the lead is determined by 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.

6.1 Human Annotated Dataset

We selected a total of 1,000 articles and split them into two sets. For the first set of 400 articles, the authors of the paper annotated the content-dense labels and provided a real-value score for the domain-dependent content-density of each text. Then a second set of 600 articles was selected and annotated on Amazon Mechanic Turk (AMT). All annotated articles were randomly picked from the NYT data and did not appear in the training data for the classifiers that we evaluate here.

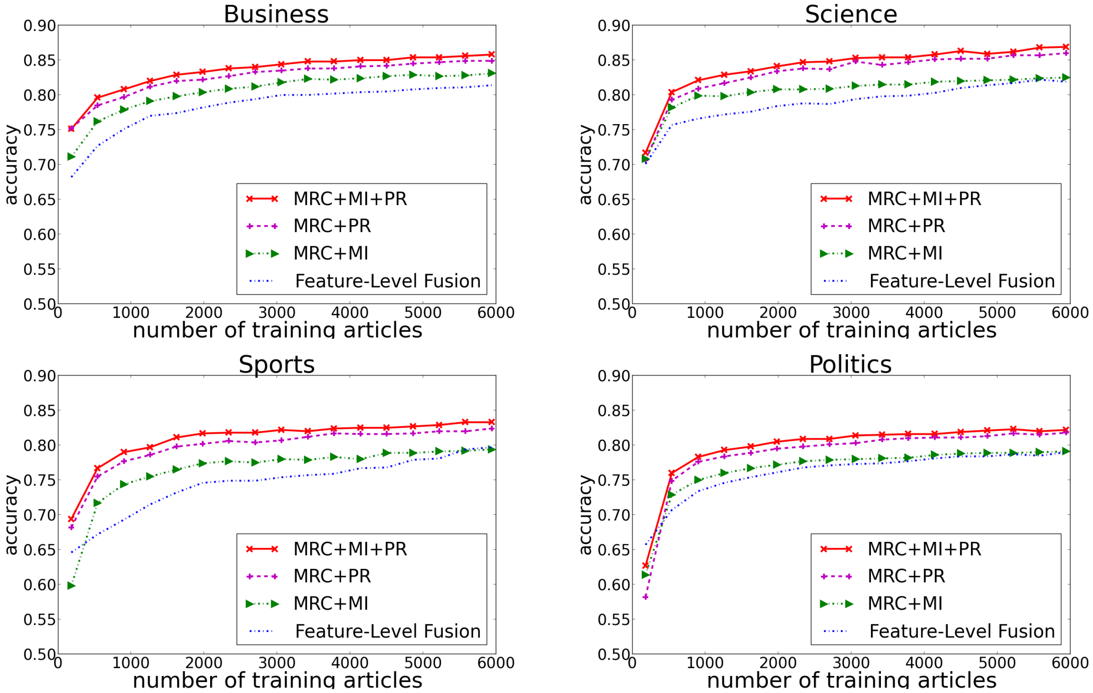


Figure 3: Accuracy by changing size of training set for the four genres: [**Top Left**] Business, [**Top Right**] Science, [**Bottom Left**] Sports, [**Bottom Right**] Politics

6.1.1 Basic Set

In the basic human annotation set, the authors of the paper annotated 400 NYT articles, 100 from each domain, with judgements of their perceived informativeness. Similar to prior work on grammatically judgements (Bard, Robertson, & Sorace, 1996), the annotation was done with respect to a reference lead that fell around the middle of the content-dense spectrum. Leads were labeled by domain: the question was if a specific article from domain D is content-dense compared to the reference lead for that domain. All 100 leads from the same domain were grouped together and displayed in random order, with the annotators seeing leads only from the same domain until they completed the annotation for that domain. The reference lead in each case was drawn from the respective domain. The annotator gave both a categorical label for the lead (less content-dense or more content-dense than the reference) and a real value score (ranging between 0 to 100) via a sliding bar. The categorical labels were used to test the binary classifiers. The real-valued annotations were used to compute correlations with classification scores produced by the classifier.

Inter-Annotator Agreement All 400 test leads were annotated as being content-dense or not and with a real-value indicator of the extent to which they are content-dense. Table 7 shows the percent agreement between the two annotators on the binary level task, as well as the correlation of the real-value annotation. For the binary annotations we also report the Kappa statistic.

Table 7: Inter-annotator agreement on manual annotations. Percent agreement is computed on the binary annotation, correlation is computed on the real-value degree of content-density of the leads. All correlations are highly significant, with $p < 0.001$.

	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

As table 7 shows, the agreement for all domains is considerably high but not perfect. Agreement is highest—almost 80%— for the politics domain. The agreement is lowest in the business domain, 70%. The correlations of content-density scores exceed 0.5 and are highly significant ($p < 0.001$) for all domains. The high correlations of real-valued scores, especially for the politics and business domains, suggest that the task may be more amenable to annotation and automation as a real-value prediction task rather than as a binary distinction.

Kappa however is relatively low, indicating that the annotation task is rather difficult. To refine our instructions for annotation, we adjudicated all leads for which there was no initial agreement on the label. Both authors sat together, reading the reference lead and each of the leads to be annotated, discussing the reasons why the lead should be labeled content-dense or not. In many cases, the final decision was made by taking into account the domain from which the lead was drawn (i.e. "there isn't much important information in a sports lead, but it could be considered content-dense in the context of sports news reporting"), as well as the reference lead for the specific genre (i.e. "the lead is not that content-dense but appears to contain more important facts or reports the news in a more direct style than the reference lead"). We study further the way domain and perception of content density interact in the next section, where independent annotators rated content-density both in in-domain and in domain-independent general settings.[7]

Below is an example on whose label the authors initially disagreed. In this lead, the first paragraph is non-informative and the second paragraph is informative, providing partial justification for either overall label.

[Example of labelling disagreement] *Many elderly people are already distressed by the increasing numbers of drugs they are taking, including painkillers and heart medication. Now, those who are also battling depression may be wondering where it all will end.*

Last week, researchers at the University of Pittsburgh presented findings from a large government-financed study suggesting that antidepressants are more effective in warding off a recurrence of late-life depression than periodic sessions of interpersonal therapy, a standardized form of talk treatment.

[7] As we will shortly see, the classifier is impressively accurate on instances in which the annotators agreed in their initial annotation and quite poor on the leads that required adjudication. These findings suggest that in future work in may be beneficial to develop a classifier for sentence-level prediction (Yang, Bao, & Nenkova, 2017) of content-density, which would be helpful for characterizing leads that mix informative and entertaining sentences. Another clear alternative is to develop a classifier to predict that a text is ambiguous in terms of its content-density status.

6.1.2 AMT Annotation Set

We also compiled a second set of 600 NYT articles, 150 for each domain. In an attempt to provide more guidance to the annotators, we gave four reference leads for each domain, two as examples of prototypical content-dense leads and two as example of leads that are clearly not content dense. The reference leads for each domain are shown in Appendix A. The annotators saw the four prototypical leads, as well as a group of leads that they had to annotate. They provided both a categorical label for each target lead (content-dense or not) and a real value score for the degree to which it can be considered content-dense (range between 0 and 100).

The annotation was partitioned into groups of five leads—an annotator had to label at least five leads and then request more data for annotation, in groups of five. To embed some quality control, one of the five leads in each group is a lead from the dataset annotated by the authors, for which they agreed in independent annotation before the adjudication step. This data allowed us to asses the quality of annotations after problematic annotators were filtered out.

Here we also study the differences in how the content-density of a text would be perceived in-domain and in general setting. For each lead text, two tasks were published separately for labeling content-density in-domain or in general. For the in-domain task, annotators are given domain information (i.e. "Here are articles drawn from the Sports section of a newspaper...") and the reference leads are selected from that domain. In the general task, workers are not told the domain of the lead and the reference leads were selected without regard to domain.[8]

Ten annotators annotated each lead in each of the two conditions.

We use two rules to filter out unqualified annotators. We filtered out all annotations by annotators who annotated too quickly or were inconsistent. The first rule is that annotator's average annotation time per task should be longer than 40 seconds. For reference, the average annotation time per task among all annotators is around two minutes. The second rule is that labeled category and score should be consistent for each lead text. If an annotator labels a lead as content-dense but gives a very low content-dense score or vice versa, we know something in their understanding of the task is amiss.

There are on average 8 annotators for each item after filtering out unqualified words. For each lead, we use the majority category as the final category label and the average score as the final score label. If there is a tie for a lead, we label it content-dense.

Table 8 shows the agreements and kappas between the majority label from AMT workers and the authors' agreed labels. We compute these only for the in-domain labels because our initial annotation was domain dependent.

Agreement for the business and sports category is high but only moderate for science and politics. We are unsure about the exact reasons why this is the case.

AMT workers annotated leads in two conditions: in-domain, where the judgements were specific to the domain from which the lead was drawn and general (domain-independent), where a domain was not specified and text from all four domains were randomly mixed in the annotation tasks. Table 9 shows the number of content-dense leads for each domain for both

[8] The content-dense example was from Business and Science, and the non content-dense from Business and Politics.

Table 8: Agreement of embedded baseline leads between AMT workers and authors of the paper.

	Agreement(%)	Kappa
Business	92.1	0.841
Science	86.8	0.622
Sports	97.3	0.947
Politics	79.0	0.574

Table 9: Number (and percentage) of content-dense leads annotated by AMT workers for each domain. The same data is annotated with respect to in-domain and general criteria and the statistics for each condition are shown in the first and last column respectively. The two middle columns show the number of leads that changed labels from content-dense(CD) to non content-dense(Non-CD) or vice versa between the in-domain and general condition, broken down according to the direction of the change.

	In-Domain	Label Changes		General
		CD → Non-CD	Non-CD → CD	
Business	93 (62.0%)	8	11	96 (66.0%)
Science	64 (42.7%)	16	25	73 (48.7%)
Sports	76 (51.1%)	38	2	40 (26.7%)
Politics	72 (48.0%)	2	53	123 (82.0%)
Overall	305 (50.8%)	64	91	332 (55.3%)

conditions, along with the number of leads whose labels changed across conditions. The first and the forth column correspond respectively to the number (percentage) of content-dense leads among all in-domain and general labels for the same data. The second and third columns show the number of labels that changed their labels from content-dense (CD) to non content-dense (Non-CD) or vice versa, between the domain-dependent and the domain-independent labelling.

Clearly, the domain context plays a large role in the perception of content density. The change is most clear for the politics and sports domain: in the domain-independent labeling a large number of sports leads, which appeared content-dense for their domain, are considered non content-dense in general. For sports, during in-domain annotation we have about half of the leads marked as content-dense, while just under 30% of the same leads are marked as content-dense in domain independent annotation. Similarly many of the politics leads considered non content-dense for the standards of the politics domain are considered as such in the domain-independent setting. There are virtually no changes in label in the opposite direction, which conforms to our expectations and provides an additional confirmation of the reasonable quality of the crowdsourced annotations.

The politics domain appears most stable, with very similar percentage of leads judged as content-dense in in-domain and general annotation. We also get some additional evidence that this domain is harder to annotate, possibly because leads there often mix both direct facts and non-literal content. We discussed this trend in our analysis of the author annotation of the domain. In the business domain, the ratio of leads that changed labels between the in-domain and general setting is closest to 1, showing least bias in perception. This is in stark contrast with politics for example, which is considered more content dense in general, attested by both the number of leads that changed label and the percentage of leads in the general setting (82%).

Overall the in-domain annotators have a more balanced number of content-dense and non content dense labels.

6.2 Are Leads Informative?

In automatic summarization research, the article leads are generally considered to be informative, or content-dense. The beginning of the article is known to be a strong summary baseline (Mani, Klein, House, Hirschman, Firmin, & Sundheim, 2002; Nenkova, 2005) and many features for identifying important content in articles are based on overlap with the opening paragraph. Our annotations allow us to directly examine to what extent this general intuition holds across domains of journalistic writing in the New York Times.

Table 9 shows the number of leads in each domain labeled as content-dense in the manually annotated dataset described above. It is clear that the prevailing assumption that the lead of the articles is always content-dense is not supported 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 system tend to be performed and which focus on specific current events, are indeed content-dense. More than 60% of leads in this domain are labeled as content-dense in the authors' annotation. The trend is similar in the AMT annotations.

Conforming to intuition, the second largest proportion of content-dense leads is in the business domain. There the articles are often triggered by current events but here is more analysis, humor and creativity. In these leads important information can often be inferred but is not directly stated in factual form. Business leads also tend to have the same labels, regardless of whether they are annotated with respect to the domain standard or in general. For the business domain, only 19 out of 150 labels changed across conditions (cf. first line in Table 9), which corresponds to at least half the rate of label change for any of the other domains.

In sports the factual information in the lead that has to be conveyed is not much and it is embellished and presented in a verbose and entertaining manner. Particularly AMT annotators consider less than a third of the sports leads to be content-dense across domains. In the science journalism section many leads only establish a general topic or an issue, or include a human interest story. Overall there is only a small partition of science leads labeled as content-dense.

The perception of content density is certainly influenced by the context of the domain. There are 55 politics leads that changed labels from in-domain to the general condition, and 53 of them are changed from non content-dense to content-dense, indicating that in that setting annotators followed their domain bias in deciding the label. Similarly 38 sports in-domain-content-dense leads are non content-dense across domains, but only 2 leads changed in the opposite direction.

These findings have two important implications for language processing applications and summarization in particular.

It is unrealistic to expect that all newspaper text has high informational value. Finding valuable content has been addressed as a standalone problem in social media (Becker, Naaman, & Gravano, 2011) and user generated data (Agichtein, Castillo, Donato, Gionis, & Mishne, 2008) but generally has been ignored in news analysis.

In addition, our analysis casts doubt on the practicality of requiring summarization systems to produce summaries of fixed length. Many of the articles with leads that are not content-dense do not discuss even in the body of the article an event readers would consider important. An appropriate summary should simply indicate this, or a summary should not be even attempted. Automatic systems are anyhow not particularly good at summarizing articles that deal with opinion or discussion rather than a specific event (Nenkova & Louis, 2008). In information access applications, tagging the genre of the article as event-centered or not (similar to earlier work in distinguishing opinion pieces from factual reporting, see Yu & Hatzivassiloglou, 2003) may be most helpful, with preview snippet summaries produced only for the event-centered articles.

6.3 Classifier Evaluation

Here we evaluate the combined two-layer classifier trained on heuristically labeled data on the manual annotations. Note that the manual annotated leads are used for evaluation only, no additional training is performed at this stage.

Following the assumption prevailing in summarization research that the lead of the article is always content-dense, the first baseline (Baseline-1) always considers the lead of the article content-dense.

The second baseline (Baseline-2) is established based on the length of the entire news article, not only the lead. The intuition is that longer articles may have uninformative leads designed to draw the reader into the subject while short articles need to start out with a more focused presentation of the event so are likely to have an content-dense lead. We train a L2-regularized logistic regression model based on this single feature. As table 10 shows, the single feature classifier achieve reasonable accuracy of 68% for the science domain.

6.3.1 Classification Results on the Basic Set

Table 10 shows the results from applying the domain-dependent and the general domain-independent models on the basic human annotation set. Accuracies computed against each of the two individual annotators is shown in the last two columns. Sports and politics domains have higher prediction accuracies on the data labeled by the first annotator, and business and science domains have higher prediction accuracies for the second annotator's labels. Also the prediction accuracies have smaller variances on the data labeled by the first annotator, between 78% for the politics domain and 74% for science the domain, compared with the accuracies on data labeled by the second annotator, between 87% for business and 71% for sports. Overall however the prediction accuracy on the final combined data, after disagreements have been adjudicated, is highest, demonstrating that the adjudication procedure did lead to more internally consistent labels. As in the heuristically labeled data, recognition accuracies are higher for the business and science domains (83%) and lower for the sports and politics domains (around 80%).

We also evaluate the prediction accuracy separately on the subsets of the data for which the two annotators agreed on the label in the first stage of independent annotation, corresponding to the presumably clear-cut cases, and those for which adjudication was needed. Clearly, the classifier captures characteristics of content-dense leads 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 to disagreement in annotation, it is more likely that the classifier makes more errors as well.

On the agreed subset—marked with the same label by both annotators during independent annotation—accuracies are around 90% for the business and science domains, 80% for sports and politics domains.

The classifier accuracies are much higher than the baselines for all domains.

We also calculate the precision, recall and F-score for the content-dense leads class for the combined dataset. The results are shown in Table 11. The domain model performs best in three of four genres, while the overall general model leads in politics. This finding is again aligned with what we observed on accuracy. Although both models can achieve good accuracy on sport leads, the F scores in that domain are not as good as in the other domains. Here both the domain model and overall model can achieve a very high precision but a relatively low recall.

Table 12 shows the correlations between the classification score from the final classifier and the real-value score of content-dense by the two annotators. All correlations are highly statistically significant. In line with what we have seen in the analysis of other results, the correlation is the highest for the business domain.

Similarly we compute the prediction accuracy stratified according to the classifier confidence in that prediction. Figure 4 shows the plot on all four genres. The accuracy of high confidence predictions is much higher than the overall accuracy. The ”article length” baseline, however, has lower accuracy in its high confidence predictions.

6.3.2 Classification Results on the AMT Annotations

Table 13 shows the accuracy and F-score of the domain-dependent and the general domain-independent models on the AMT annotations. As in previous tables, row 1 and 2 represent the results from domain models and the domain-independent models respectively. Rows 3 to 4 show results for the two baselines. Our classifiers outperform the baselines by a large margin except for politics in the domain-independent labels, where the baseline that considers all leads to be content dense works best. Overall however, the results show that the baseline of assuming all leads are content-dense performs poorly and the proposed approaches significantly improve the accuracies.

Comparing the accuracies of prediction for data drawn from the same newspaper section, it is evident that business and science have the most stable prediction and the accuracy of the domain-dependent and the domain-independent classifiers does not differ much on these subsets of the test data. The classifier trained on domain-independent labels achieves 78.0% accuracy on the domain-specific labels, in which the annotators were explicitly told

Table 10: Binary classification accuracies(%) on basic human annotated datasets for models trained on heuristically labeled data.

Business	Combined	Agreed	Adjudicated	Anno_1	Anno_2
Domain model	**83**	**94.3**	**56.7**	**75**	**87**
Overall model	79	91.4	50.0	**75**	83
Baseline-1	53	52.8	53.3	47	57
Baseline-2	60	65.7	46.7	58	64
Science	Combined	Agreed	Adjudicated	Anno_1	Anno_2
Domain model	**83**	**89.2**	**65.4**	**77**	80
Overall model	81	**89.2**	57.7	71	**86**
Baseline-1	37	31.1	53.8	45	27
Baseline-2	68	69	**65.4**	62	65
Sports	Combined	Agreed	Adjudicated	Anno_1	Anno_2
Domain model	**78**	**80.8**	70.3	**74**	**71**
Overall model	75	75.3	**74.1**	69	68
Baseline-1	49	46.5	55.6	45	50
Baseline-2	65	70	51.9	63	66
Politics	Combined	Agreed	Adjudicated	Anno_1	Anno_2
Domain model	78	**83.3**	59.1	**78**	74
Overall model	**80**	**83.3**	**68.2**	76	**76**
Baseline-1	61	60.3	63.6	55	61
Baseline-2	51	55	36.4	55	53

Table 11: [P]recision, [R]ecall and [F]score (%) on basic human annotated datasets for models trained on heuristically labeled data. [D]omain model, [O]verall model, and [B]aseline-2.

	Business			Science			Sports			Politics		
	P	R	F	P	R	F	P	R	F	P	R	F
D	**81**	**88.7**	**84.7**	**95.7**	**57.9**	**72.1**	**90.9**	**50**	**64.5**	**95.4**	67.7	79.2
O	76.8	84.3	80.4	91.3	55.2	68.8	86.3	**50**	63.3	87.3	**78.7**	**82.7**
B-2	61	67.9	64.3	61.5	42.1	50	66.7	57.1	61.5	63	47.5	54.2

Table 12: Correlation between predicted probabilities and human annotated scores. All correlations are highly significant with $p < 0.001$.

	Annotator_1		**Annotator_2**	
	Domain Models	Overall Models	Domain Models	Overall Models
Business	0.621	**0.647**	0.797	**0.810**
Science	**0.575**	0.546	0.711	**0.758**
Sports	**0.590**	0.575	**0.588**	0.582
Politics	**0.658**	0.629	**0.609**	0.592

the news section from which the article was drawn and used this information in judging if the lead is content-dense or not. This accuracy is less than 2% lower than the prediction on

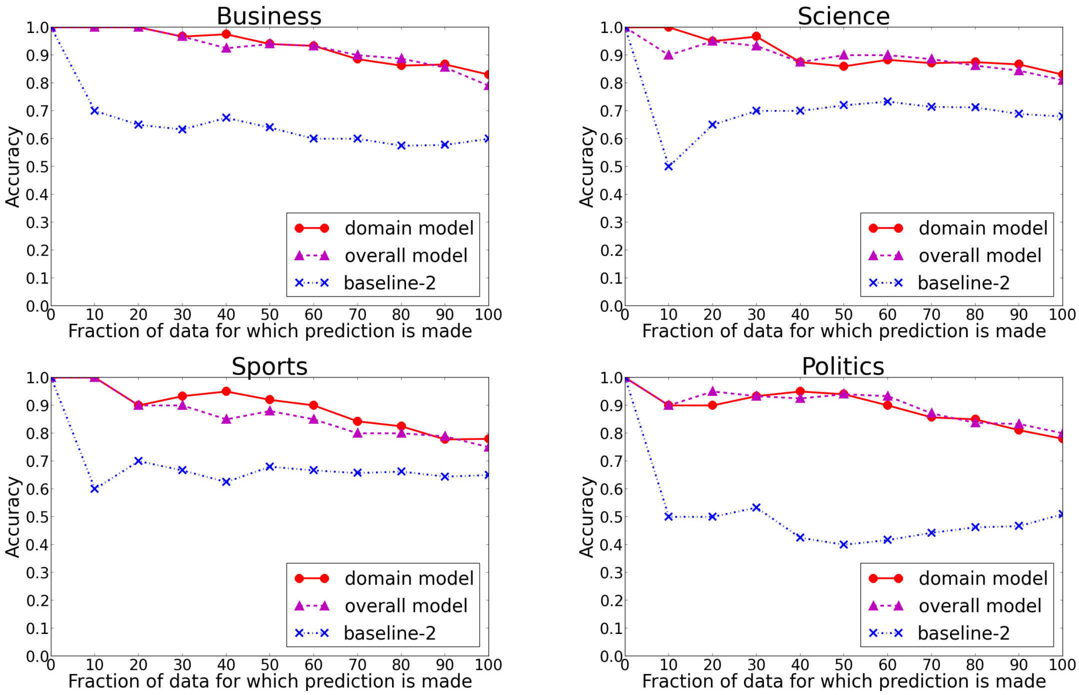


Figure 4: Prediction accuracy based on probability ranking on basic human annotation set. [**Top Left**] Business, [**Top Right**] Science, [**Bottom Left**] Sports, [**Bottom Right**] Politics

domain-independent labels. Similarly in the science domain the difference in performance for the two types of labels is as low as 2.0%. In stark contrast, there is large difference in performance for the in-domain and domain-independent labels for sports and politics, where the difference between the two reaches 10%.

The crowdsourced annotations were performed both in-domain (judging the content density with respect to the expectation for the given domain, politics for example) and in domain independent setting. Here domain models are on average worse than the domain independent models. For the sports domain, training a domain-specific classifier helps most in improving the detection of content-dense sports leads but for the other domains the advantage is less clear. This result is reassuring. If the domain models were clearly superior, one would have needed accurate domain predictors for practical applications. The analysis presented here demonstrates that a domain-independent classifier may be sufficient for many applications.

The accuracies of the domain models drop considerably compared to their respective accuracies on the author-annotated set. For example, there is around 8.0% drop in the politics domain. There are several possible reasons for this difference. The articles in the initial set that the authors annotated were selected only from the articles published in 2005 and 2006 while the AMT set is selected from the entire NYT dataset from 1987 to 2007. The annotation instructions also differed for the two sets. The AMT annotators were presented with prototypical content-dense and non content-dense leads as references, while the authors had only one lead in the middle of the range of content-density as reference. Finally, the general domain-independent classifier on average works best, predicting both the in-domain and general labels in the test set better than the domain-dependent classifiers. This trend indicates that AMT workers were likely more influenced by general domain expectations when labeling the data. It is plausible that domain-dependent annotation requires more detailed instructions that are not as readily passed on in the crowdsourced setting.

Table 13: Binary classification results on AMT annotated datasets for models trained on heuristically labeled data: [A]ccuracies(%) and [F]scores(%)

In-domain	Business		Science		Sports		Politics		Average	
	A	F	A	F	A	F	A	F	A	F
Domain model	76.0	78.3	72.7	70.9	**73.3**	**69.5**	**70.7**	**73.5**	73.2	73.5
General model	**78.0**	**80.4**	**76.7**	**74.8**	70.0	63.1	70.0	73.3	**73.7**	**73.9**
Baseline-1	62.0	–	42.7	–	51.1	–	48.0	–	55.0	–
Baseline-2	58.0	62.7	64.7	40.0	62.7	51.2	52.7	48.5	59.5	61.4
Domain indep.	Business		Science		Sports		Politics		Average	
	A	F	A	F	A	F	A	F	A	F
Domain model	**79.3**	**81.4**	73.3	74.0	80.0	62.8	76.7	83.8	77.3	75.5
General model	77.3	80.4	**78.7**	**78.4**	**82.0**	**67.6**	77.3	**84.5**	**78.8**	**77.7**
Baseline-1	66.0	–	48.7	–	26.7	–	**82.0**	–	55.9	–
Baseline-2	62.7	68.6	66.7	62.8	68.0	53.3	50.7	61.1	62.0	61.7

We further compute the correlations coefficients between predicted probabilities and average scores annotated by AMT workers. The results are shown in Table 14. Domain models have better correlations than general models in three of domains for domain dependent (in-domain) labels, but with small absolute difference in correlation. The domain-independent models are much better in predicting content-dense in the general, domain-independent condition. All correlations are highly significant, ranging from 0.577 to 0.661 against in-domain labels and from 0.602 to 0.730 against domain-independent labels. As in the binary prediction task, the domain-independent label appear to be easier for the system to predict. The correlation coefficients are in line with our intuition and much closer to the numbers we have seen based on the basic author-annotated set (shown in Table 12). This trend implies that predicting content-density in terms of real-value scores may be more suitable for this task.

Table 14: Correlation between predicted probabilities and average scores annotated by AMT workers in the domain specific and general condition. All correlations are highly significant with $p < 0.001$.

	In-Domain Labels		**Domain Independent Labels**	
	Domain Models	General Model	Domain Models	General Model
Business	0.602	**0.614**	0.713	**0.730**
Science	**0.661**	0.646	0.652	**0.690**
Sports	**0.600**	0.577	**0.619**	0.602
Politics	**0.616**	0.615	0.652	**0.668**

For the AMT annotated test set, we also compute the prediction accuracy stratified according to percentiles of data ranked by the classifier confidence in that prediction. Figures 5 and 6 show the plots on all four domains for the two types of annotated labels (domain-specific or domain-independent). Again, the accuracy of high confidence predictions is much higher than the overall accuracy. The article length baseline, however, has much lower accuracies.

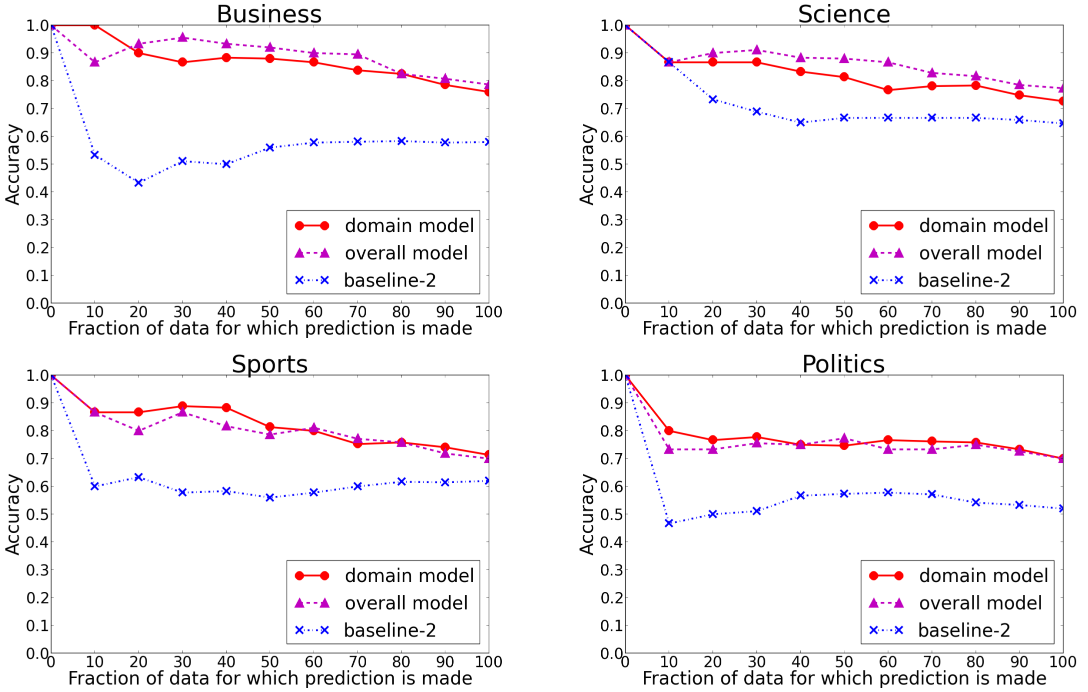


Figure 5: Prediction accuracy based on probability ranking on AMT annotated data. The x axis represents the percentile of data used to calculate the accuracy according to the predication confidence. (In-domain): **[Top Left]** Business, **[Top Right]** Science, **[Bottom Left]** Sports, **[Bottom Right]** Politics

7. Recognizing Better Summaries

So far we have demonstrated that detector of content-density can be developed using heuristically labeled data and that it can achieve respectable accuracy in intrinsic evaluation on human-labeled leads. Ultimately however the goal would be to integare the content-density prediction in information seeking applications such as summarization and news browsing. Testing the impact of the content-density prediction in such extrinsic evaluations will be the main focus of future work.

Here, however, we show a feasibility study to verify the potential for development of more informed summarization methods that exploit the concept of content density. Specifically, we demonstrate that the content-density detector is able to recognize when an automatic summary of a single news article is better than the lead of the article. This is an important open problem in summarization, where the lead paragraph baseline is very strong and few

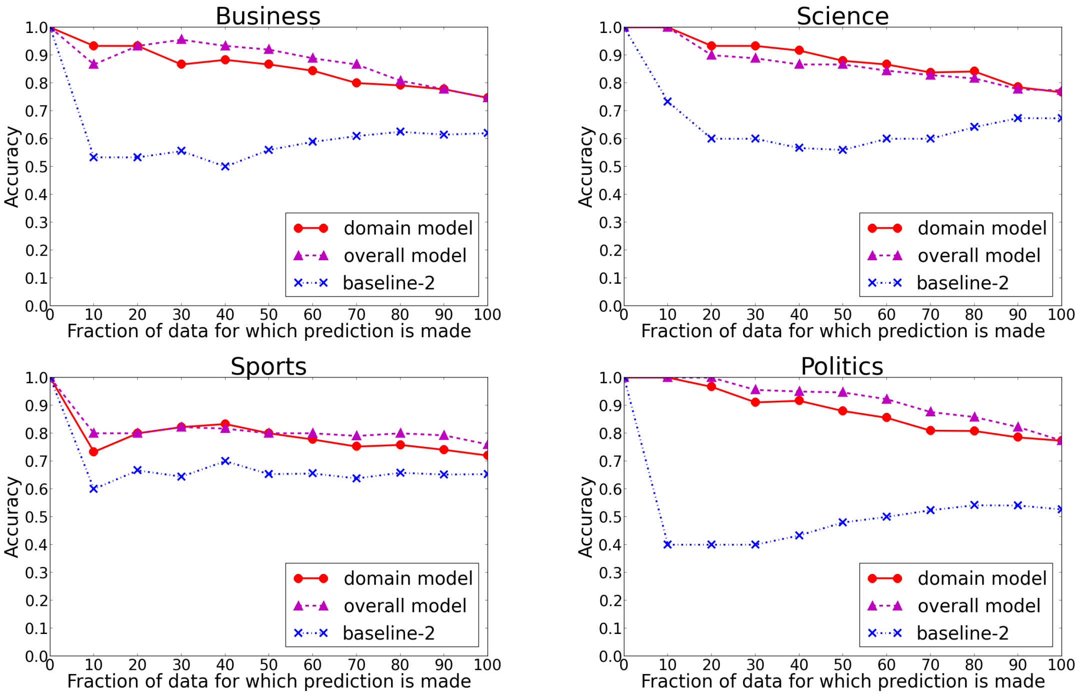


Figure 6: Prediction accuracy based on probability ranking on AMT annotated data. The x axis represents the percentile of data used to calculate the accuracy according to the predication confidence. (Domain independent): [**Top Left**] Business, [**Top Right**] Science, [**Bottom Left**] Sports, [**Bottom Right**] Politics

systems outperform it (Over, Dang, & Harman, 2007). Moreover, all of the proof-of-concept experiments from the previous section were performed on data drawn from the NYT. We would like to verify that the usefulness of the prediction remains when data from other sources is considered.

Motivated by these goals, we perform our experiments on detecting if a machine summary is more informative than the lead on two datasets: NYT and on data from the Document Understanding Conference, which has data from a variety of sources.

We randomly selected 400 articles with manual summaries from the NYT, 100 for each genre. We generated automatic summaries for the article using two systems. The first system, LeadSumm, is the strong lead baseline which picks the first 100 words as the summary. The second system is IcsiSumm (Gillick & Favre, 2009), which is one of the state-of-the-art multi-document summarization systems (Gillick, Riedhammer, Favre, & Hakkani-Tur, 2009; Berg-Kirkpatrick, Gillick, & Klein, 2011).

We performed human evaluation to determine which of the two summaries in the pair (LeadSumm and IcsiSumm) is better. We asked annotators to first read the manual summary from NYT, then read the two summaries generated by the systems. Then we asked the annotators to indicate which of the two system summaries covers better the information expressed in the NYT goldstandrd. They were also provided with an option to indicate that the two system summaries cover the information expressed in the goldstandard equally well.

The flow between the sentences in the LeadSumm summaries is better than those in the automatic summaries because this is a snippet of professionally written discourse. Here our goal is to study the content-density of the two summaries, independently of the linguistic quality of the summary which we know favors the lead system. For this reason, we randomized the order of the sentences in both of the LeadSumm and IcsiSumm summaries. The order of presenting the LeadSumm and IcsiSumm to annotators is also randomized during judgement collection.

The tasks are published on Amazon Mechanical Turk (AMT) and each task is assigned to 10 annotators, with one summary per task/HIT. IcsiSumm generated empty summaries for 77 out of the 400 randomly selected articles and two of its summaries were identical to the lead baseline. We removed those tasks so there are total 323 tasks are published. The majority vote of the 10 annotators is used as the final label. The human annotations are used as ground truth in the following steps.

Next we apply the content-dense detector on the generated IcsiSumm and LeadSumm to get content-dense probability scores for each. The summary with higher content-dense score is predicted to be the better summary. As expected from prior manual shared task evaluations, LeadSumm is better then IcsiSumm for most of the articles.

The confidence in the prediction that one summary is better than another is controlled by the content-dense score difference. The larger the difference between the content-density scores of the IcsiSumm and LeadSumm is, the more confident we can be that the summary is indeed better. We track how the summarization performance varies with the difference in content-density scores. In cases when the difference between the two content-density scores is lower than a set value, we consider that the lead summary is better. By defining the $score\_difference = score_{IcsiSumm} - score_{LeadSumm}$, we cutoff the evaluations samples by $score\_difference$ compute the metrics for different cutoff levels.

Table 15 shows the results of detecting when IcsiSumm produces better summaries than the lead baseline on the NYT articles. This is equivalent to a combination system which uses lead summaries unless it is confident that the automatic summary is better, in which case it uses the IcsiSumm summary. The first column represents the cutoff value and the second column shows the number of total samples within this cutoff. Column 3 to 5 are the statistics of human judgements. The last column shows the number (percentage) of correctly predicted samples of the combination system. Each row shows the results of the system with a cutoff value. The last row shows the statistics for the entire dataset and two baselines, one that picks ICSISumm summaries only and another that picks lead summaries only. The lead was better for 59% of the test articles; the IcsiSumm produced the more informative summary for 34% of the test articles. The two summaries were considered equally informative in the rest of the cases. Clearly, as expected from past DUC evaluations, the lead baseline summary is better than the automatic summarizer. However even an extractive summarizer can significantly outperform the lead baseline if we had a reliable way in which to predict when an alternative summary would be more informative; this could improve one out of each three summaries produced by the summarizer.

Last column shows the performance of a combination system using content-density scores to decide which of the two available summaries is superior. Whenever at least some threshold is used to decide when the automatic summary is better, the combination system's output is preferred by the assessors considerably more often than the output for the lead baseline.

Particularly for thresholds between 0.1 and 0.4, the output of the combination system is preferred between 64 and 66% of the time, compared to the 58.5% for the lead baseline. These improvements are statistically significant (with $p < 0.05$) according to a binomial test with expected probability of producing better summary of 0.585, corresponding to the human preference for the baseline lead summaries.

Table 15: Performance of combination system with different cutoffs on NYT articles. The last column shows the number (percentage) of correct predicted samples of the combination system.

	cutoff	# of samples	Human Judgement			
			Icsisumm	Tie	Leadsumm	Combination System
NYT	0.5	18	14	0	4	199 (61.6%)
	0.4	29	23	1	5	207 (64.1%)
	0.3	42	32	2	8	213 (65.9%)
	0.2	54	39	2	13	215 (66.6%)
	0.1	79	47	4	28	208 (64.4%)
	0	179	78	7	94	173 (53.6%)
	All	323	109 (33.7%)	25	189 (58.5%)	N/A

Next we verify that the proposed model works with reasonable accuracy on sources other than the NYT. We run the same LeadSumm and IcsiSumm systems on the DUC dataset (Over et al., 2007). We only perform the experiments on the data from DUC2002, which is the last year NIST provides single-document human summaries.

There are total of 533 articles from various sources in DUC2002, including Associated Press (AP), Wall Street Journal (WSJ), Los Angeles Magazine (LA), FT Magazine (FT), San Jose Mercury News (SJMN) and Foreign Broadcast Information Service Daily Reports (FBIS). AP is a newswirse service, providing high-quality news reporting used by many media outlets. By the nature of newswire services, the AP articles (and leads) are expected to be contain a larger propartition of content-dense texts. The other article sources are drawn from newspapers, so are much more likely to include leads that are not content-dense.

Again, we filter out articles for which IcsiSumm summary could not be generated or for which the automatic and lead summary were identical. Again, we exclude from considerations articles for which IcsiSumm did not generate a summary or for which IcsiSumm produced a summary consisting of the lead of the article. After filtering these, we obtain 493 articles for the evaluation.

As with the NYT experiment, we use AMT to obtain judgements about which of the two summaries of the article is better. All the annotation settings are exactly the same with the NYT annotations described above.

Again, we then apply content-dense detector on the generated IcsiSumm and LeadSumm to detect which summary is better based on their content-dense scores. Table 16 shows the results of detecting better IcsiSumm on the DUC2002 articles.

The judgements on data drawn from sources different from the NYT allows us to get a sense about the extent to which the content-density detector we developed to the news genre in general rather than specific to the NYT. The observation that bears special mention is

Table 16: Performance of combination system with different cutoffs on DUC2002 articles. The last column shows the number (percentage) of correct predicted samples of the combination system.

	cutoff	# of samples	Human Judgement			
			Icsisumm	Tie	Leadsumm	Combination System
DUC AP	0.5	5	4	0	1	246 (78.6%)
	0.4	13	9	0	4	248 (79.2%)
	0.3	22	11	1	10	244 (78.0%)
	0.2	38	15	1	22	236 (75.4%)
	0.1	72	21	2	49	215 (68.7%)
	0	117	30	5	82	191 (61.0%)
	All	313	51 (16.3%)	19	243 (77.6%)	N/A
DUC Other	0.5	4	3	0	1	134 (74.4%)
	0.4	6	5	0	1	135 (75.0%)
	0.3	8	6	1	1	136 (75.6%)
	0.2	9	6	2	1	136 (75.6%)
	0.1	18	9	4	5	135 (75.0%)
	0	48	19	5	27	123 (68.3%)
	All	180	41 (22.7%)	8	131 (72.8%)	N/A

that here, the percentage of articles for which the IcsiSumm is able to produce more informative summaries than the lead summarizer is considerably smaller than in the randomly selected sample of NYT articles. For the AP articles, the automatic summaries are judged as better than the lead baseline for 16% of the test articles. This conforms with expectations that the AP articles and leads will be overall more content-dense than regular newspaper sources. For the other sources (newspaper), the percentage of automatic summaries that are better than the lead is 23%. For comparison, in the NYT sample the system produced a more informative summary in 34% of the cases. This larger percentage may reflect the style of the New York Times or the fact that the articles from NYT were randomly drawn, so covers a broader range of domains than the DUC data.

The numbers indicate that the style of AP is the most typically informational while the NYT is the most stylistically rich, with leads that are often not content-dense. If this is the case, the room for expediting news search and browsing via automatic summarization has been underestimated in DUC evaluations.

The last column shows the performance of the combination system. The performance on DUC Other is still similar to the performance on NYT. Just picking any cutoff ≥ 0.1 will lead a performance improvement. However, the combination system is only better than lead baseline for AP when the cutoff ≥ 0.3 and reaches the best performance when setting cutoff to 0.4, which is aligned with previous finding that AP is typically informational.

Overall, the combination system performs better than baseline systems when we pick the right cutoff. The choice of cutoff depends on the source types and the reason is the writing styles can be very different in different sources as discussed above. For NYT and DUC newspapers sources (excluding AP), the combination system is able the achieve better performance when setting cutoff as 0.1 and achieve highest accuracy when setting cutoff as 0.2. For the AP, however, it is much harder to find a cut off in which the combination system would outperform the lead baseline. The cutoff has to be set to 0.4 to get the best performance, so the general ability to produce a better summary with AP data is dubious.

This analysis of cut-offs at which prediction of content density would improve summarization is only a pilot analysis. Ideally, we would need sufficient data to have a dedicated development set on which to determine the cut-off and an independent test set on which to verify its utility. Such in-depth study is left for future work. Here however we note that there is a clear difference in the performance for the stylistically different newswire and newspaper and that there is sufficient evidence for potential of using the content-density stylistic distinction for improving single document summarization.

8. Conclusion and Future Work

In this paper we introduced the task of detecting content-dense news article leads. We use article/summary pairs from the NYT corpus to heuristically label a large set of articles as content-dense when the lead of the article overlaps highly with the human summary and as non content-dense when the overlap is low.

We present experiments with two lexical representations and one syntactic representation. The production rule syntactic representation is the best predictor of lead content-density among the three. The corpus-independent lexical representation from a vocabulary defined by the MRC lexicon proved to be the more useful lexical representation. We compared a feature-level combination model and a two-layer decision-level combination model. The latter performs best in all our experiments.

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

Overall domain-specific models are more accurate than in-domain labels from trained annotators. The general model trained on all data pooled together achieves better performance on crowdsourced annotations in both domain-dependent and domain independent annotation conditions. Our experiments indicate that predicting content-dense in terms of real-value scores may be more accurate and beneficial for applications than simply classifying a lead as content-dense or not.

In this work, we have established the feasibility of the task of detecting content-dense texts. We have confirmed that the automatic annotation of data captures distinctions in informativeness as perceived by people. We also show proof-of-concept experiments that show how the approach can be used to improve single-document summarization of news and the generation of summary snippets in news-browsing applications. In future work the task can be extended to more fine-grained levels, with predictions on sentence level and the predictor will be intergared in a fully functioning summarization system.

All data for the work presented in this paper and the domain-dependent and general classifiers will be made publicly with the publication of this article.

Acknowledgments

We reported our initial work on detection of information-dense text in a paper published at AAAI 2014 (Yang & Nenkova, 2014) In this manuscript we have further extended the work by using much larger samples of New York Times articles for training and by analyzing the performance of a larger number of two-layer classifiers. Here we also introduce a new collection of manually annotated test data, for both domain-dependent and general content-density of texts. We make use of this collection for detailed evaluation of the content-density classifiers. In our AAAI 2014 paper, we use the term information-dense to describe the types of text we wish to detect. Here we switch the terminology to content-dense, to avoid confusion with work in the intersection of cognitive science and computational linguistics that uses the term information density in information-theoretic sense to describe the change in surprise in the linear processing of sentences (Jaeger, 2010; Pate & Goldwater, 2015).

Appendix A. Reference Leads Used in AMT Annotations

Here we present all the references leads annotators saw in AMT human intelligent tasks (HITs).

A.1 In-Domain Reference Leads

In in-domain annotations, annotators labeled a group of five leads from same domain in each HIT. Two content-dense leads and two non content-dense leads from the same domain are displayed at the beginning.

A.1.1 Reference Leads for Business Domain

[**Content-dense Ref 1**] Securities regulators charged one of the richest men in Mexico, Ricardo B. Salinas Pliego, with fraud yesterday, in a lawsuit that seeks to have him barred as a director or officer of any company whose shares trade on an American exchange.

The Securities and Exchange Commission also sought to have Mr. Salinas Pliego, the chairman of TV Azteca, the second-biggest Spanish-language broadcaster, give up more than $110 million he made from trading in the company's stock and debt.

[**Content-dense Ref 2**] In a rare move, Microsoft said yesterday that it had agreed to pay a percentage of the sales of its new portable media player to the Universal Music Group.

Universal Music, a unit of Vivendi, will receive a royalty on the Zune player in exchange for licensing its recordings for Microsoft's new digital music service, the companies said.

[**Non content-dense Ref 1**] LOOKING for some thong underwear or perhaps a leather jacket and don't know where to find them? Try logging on to a restaurant Web site.

Small restaurateurs are increasingly using the Internet to sell goods that go far beyond the usual array of branded T-shirts and hats, in hopes of not just building the bottom line, but also cultivating possible new markets for expansion.

[**Non content-dense Ref 2**] "WHAT stresses me most," the chief executive of Novartis, Daniel L. Vasella, said, "is that we are getting new regulations from abroad without any consultation.“

This has been the World Economic Forum that the United States government largely passed by. In a world that both respects and fears American power, there is worry that the United States does not care what others think.

A.1.2 Reference Leads for Science Domain

[**Content-dense Ref 1**] Scientists have decoded the chimp genome and compared it with that of humans, a major step toward defining what makes people human and developing a deep insight into the evolution of human sexual behavior.

The comparison pinpoints the genetic differences that have arisen in the two species since they split from a common ancestor some six million years ago.

[**Content-dense Ref 2**] A popular class of drugs for high blood pressure, ACE inhibitors, may cause birth defects if taken during the first three months of pregnancy, doctors are reporting. Pregnant women and those who are planning to become pregnant should avoid the drugs, the researchers and officials at the Food and Drug Administration warn.

ACE inhibitors have long been known to cause birth defects if taken later in pregnancy, but until now were considered safe if taken in the first trimester.

[**Non content-dense Ref 1**] To gauge the potential consumer impact of the consolidation sweeping the telephone industry, look no further than the silver-toned plastic phone gathering dust on the desk in Justin Martikovic's studio apartment.

Mr. Martikovic, 30, a junior architect who relies on a cellphone for his normal calling, says he never uses the desk phone – but he pays $360 a year to keep it hooked up.

[**Non content-dense Ref 2**] As the horror of the South Asian tsunami spread and people gathered online to discuss the disaster on sites known as Web logs, or blogs, those of a political bent naturally turned the discussion to their favorite topics.

To some in the blogosphere, it simply had to be the government's fault.

A.1.3 Reference Leads for Sports Domain

[**Content-dense Ref 1**] Ivor G. Balding, one of three British brothers who gained international fame as polo stars in the 1930's, when the sport attracted large crowds and wide press coverage, died on Thursday at his home in Camden, S.C. He was 96.

His death was announced by his family.

[**Content-dense Ref 2**] Finally, the deal is done.

Laveranues Coles, the wide receiver from the Washington Redskins, passed a physical examination by the Jets' medical staff yesterday, clearing the way for the team to reacquire him in a trade for wide receiver Santana Moss.

[**Non content-dense Ref 1**] NEARLY 36 years ago, when it was his turn to interview the prospective employee, the estimable James Reston, onetime traveling secretary for the

Cincinnati Reds but then the executive editor of this newspaper, asked how a political science major had wound up writing about sports.

I answered the question, but I have a better answer now. The political science classes prepared me for the nonsense that will pass for a hearing about steroids use in baseball next Thursday in Washington.

[**Non content-dense Ref 2**] Three years ago, as he stood in the rubble of the St. Bonaventure basketball program, Ahmad Smith had a decision to make.

One of his teammates, center Jamil Terrell, had been declared ineligible after it was learned that he had been admitted to the Franciscan university in the hills of southwestern New York with a welding certificate – and the approval of St. Bonaventure's president.

A.1.4 Reference Leads for Politics Domain

[**Content-dense Ref 1**] At least 844 American service members were killed in Iraq in 2005, nearly matching 2004's total of 848, according to information released by the United States government and a nonprofit organization that tracks casualties in Iraq.

The deaths of two Americans announced by the United States military on Friday – a marine killed by gunfire in Falluja and a soldier killed by a roadside bomb in Baghdad – brought the total killed since the war in Iraq began in March 2003 to 2,178. The total wounded since the war began is 15,955.

[**Content-dense Ref 2**] Seventeen people died in two separate violent incidents on Sunday and Monday that underscored an increasing sense of lawlessness in Mexico.

A former soldier went on a rampage in a Pacific coast town on Sunday, killing 12 people before local residents chased him down and the police shot him in the town square. Thirteen hours later, gunmen attacked gamblers at an illegal cockfight at a Guadalajara racetrack, killing 4 and wounding 27 when they tossed two grenades into the crowd.

[**Non content-dense Ref 1**] President Bush on Tuesday pressed Senate Republican leaders to continue fighting to confirm John R. Bolton as ambassador to the United Nations, even though Senator Bill Frist, the majority leader, said his options had been exhausted and some Republicans urged the appointment of Mr. Bolton when Congress recesses.

”The president made it very clear that he expects an up-or-down vote,“ Dr. Frist told reporters after meeting with the president. Back in the Capitol, he added, ”I don't want to close that door yet.“

[**Non content-dense Ref 2**] ARE things getting better or worse in Iraq? That is the basic question, on which much hinges for the United States and the world. Here are some impressionistic answers.

Just over a year ago, on my last visit to the country, I was able to drive north to Tikrit, Saddam Hussein's home town, and south to the Shiite holy city of Najaf. These were not excursions for sitting back and enjoying the scenery. But they were feasible, at high speed and with some risk.

A.2 Domain-Independent Annotation

In domain-independent annotations, annotators are given a group of five leads randomly selected from all domains. Two informative leads and two uninformative leads are given as references.

[**Content-dense Ref 1**] Securities regulators charged one of the richest men in Mexico, Ricardo B. Salinas Pliego, with fraud yesterday, in a lawsuit that seeks to have him barred as a director or officer of any company whose shares trade on an American exchange.

The Securities and Exchange Commission also sought to have Mr. Salinas Pliego, the chairman of TV Azteca, the second-biggest Spanish-language broadcaster, give up more than $110 million he made from trading in the company's stock and debt.

[**Content-dense Ref 2**] In a rare move, Microsoft said yesterday that it had agreed to pay a percentage of the sales of its new portable media player to the Universal Music Group.

Universal Music, a unit of Vivendi, will receive a royalty on the Zune player in exchange for licensing its recordings for Microsoft's new digital music service, the companies said.

[**Non content-dense Ref 1**] LOOKING for some thong underwear or perhaps a leather jacket and don't know where to find them? Try logging on to a restaurant Web site.

Small restaurateurs are increasingly using the Internet to sell goods that go far beyond the usual array of branded T-shirts and hats, in hopes of not just building the bottom line, but also cultivating possible new markets for expansion.

[**Non content-dense Ref 2**] As the horror of the South Asian tsunami spread and people gathered online to discuss the disaster on sites known as Web logs, or blogs, those of a political bent naturally turned the discussion to their favorite topics.

To some in the blogosphere, it simply had to be the government's fault.

Appendix B. Production Rules with Highest Weights

In this section we list the production rules with highest weights for each genre. We also show two example for each production rule. The examples are extracted from lead texts using Stanford CoreNLP package.

Table 17: Top 10 production rules with examples for Business

Positive Rules
$^{+}$**VP->VB NP PRT ADVP**
1) VP ->VB[scare] NP[them] PRT[away] ADVP[all over again] 2) VP ->VB[push] NP[the Czech currency] PRT[up] ADVP[sharply]
$^{+}$**VP->VBG PP S**
1) VP ->VBG[boasting] PP[on line about their incentive packages] S[to attract companies to relocate to their areas] 2) VP ->VBG[looking] PP[for facts about different regions] S[to get information that only used to be available , if at all , through the mail and in-person visits]
$^{+}$**NP->NN**
1) NP ->NN[response] 2) NP ->NN[overdrive]
$^{+}$**VP->ADJP VBG NP PP**
1) VP ->ADJP[tough] VBG[protecting] NP[American industry] PP[from unfair trading practices] 2) VP ->ADJP[sometimes heated] VBG[questioning] NP[Tuesday] PP[from members of a House subcommittee]
$^{+}$ **NP->DT NNP**
1) NP ->DT[the] NNP[I.M.F.] 2) NP ->DT[the] NNP[F.D.A.]

Negative Rules
$^{-}$ **NP->JJ CD NNS**
1) NP ->JJ[pre-April] CD[15] NNS[blues] 2) NP ->JJ[past] CD[150] NNS[degrees]
$^{-}$ **VP->VBN PP NP-TMP PP**
1) VP ->VBN[injured] PP[in a car crash in Peru , a third weathered] NP-TMP[a summer] PP[in Pakistan in brutal 117-degree heat] 2) VP ->VBN[swayed] PP[down the wet black runway at the Alexander McQueen fashion show last Thursday] NP-TMP[night] PP[to an ominous disco]
$^{-}$ **ADVP->RBR RB PP**
1) ADVP ->RBR[more] RB[often] PP[than not] 2) ADVP ->RBR[More] RB[often] PP[than not]
$^{-}$ **VP->VBZ : NP**
1) VP ->VBZ[War] :[:] NP[Has Newsweek 's Time Finally Come] 2) VP ->VBZ[is] :[:] NP[Now what]
$^{-}$ **VP->VBD ADVP NP-TMP , NP**
1) VP ->VBD[fell] ADVP[sharply] NP-TMP[yesterday] ,[,] NP[the fourth consecutive decline , as concerns about inflation and interest rates grew before today 's report on producer prices] 2) VP ->VBD[opened] ADVP[here] NP-TMP[Friday] ,[,] NP[another sign of how companies all over the world are still rushing to do business in China]

Table 18: Top 10 production rules with examples for Science

Positive Rules
⁻ **QP->JJR IN NP**
1) QP ->JJR[more] IN[than] NP[the vast majority] 2) QP ->JJR[more] IN[than] NP[a jubilant return]
⁻ **ADJP->ADJP SBAR**
1) ADJP ->ADJP[less likely than others to have children , and those who do give birth run an increased risk of bearing a child with the same birth defect] SBAR[that they themselves have] 2) ADJP ->ADJP[far less successful] SBAR[than expected]
⁻ **NP->DT NNP NNS NN**
1) NP ->DT[A] NNP[Federal] NNS[appeals] NN[court] 2) NP ->DT[a] NNP[Texas] NNS[appeals] NN[court]
⁻ **S->FRAG NP VP .**
1) S ->FRAG[In] NP[Old] VP[Souls : The Scientific Evidence For Past Lives , ” -LRB- Simon Schuster , 1999 -RRB- Tom Shroder , a Washington Post editor , reviews the 80-year-old clinical psychiatrist ’s research on reincarnation and finds it hard to refute] .[.] 2) S ->FRAG[Tonight , when] NP[Live From Lincoln Center ”] VP[broadcasts a concert by the New York Philharmonic on PBS stations across the country , the announcer will not be saying anything about the personal story of the bass-baritone Thomas Quasthoff , who will sing four concert arias by Mozart] .[.]
⁻ **NP->PRP$ NNS NN**
1) NP ->PRP$[their] NNS[doctors] NN[charge] 2) NP ->PRP$[their] NNS[employees] NN[home]

Negative Rules
⁻ **VP->VBG NP PP PP**
1) VP ->VBG[ordering] NP[a cup of coffee] PP[at Starbucks] PP[into an Olympic challenge] 2) VP ->VBG[taking] NP[a crack] PP[at his plays] PP[in the form of faithful revivals or loose interpretations]
⁻ **VP->VB NP ADVP , SBAR**
1) VP ->VB[get] NP[both his legs] ADVP[amputated] ,[,] SBAR[even though they had been perfectly healthy] 2) VP ->VB[use] NP[her niece ’s card] ADVP[here] ,[,] SBAR[since she does n’t live in Westchester]
⁻ **VP->VBN NP , ADVP PP**
1) VP ->VBN[triggered] NP[copycats] ,[,] ADVP[sometimes] PP[by the dozens] 2) VP ->VBN[been] NP[7,000 cases of leprosy in this country over the previous three years] ,[,] ADVP[far more than] PP[in the past]
⁻ **NP->NP , NP CC NP**
1) NP ->NP[social X-rays] ,[,] NP[those rail-thin women who had attained the exalted status that comes from being married to a Master-of-the-Universe investment banker] CC[or] NP[lawyer] 2) NP ->NP[your new book] ,[,] NP[Evolution ’s Rainbow : Diversity , Gender and Sexuality in Nature] CC[and] NP[People]
⁻ **SBAR->SBAR , RB SBAR**
1) SBAR ->SBAR[all about whom we could persuade to hire us] ,[,] RB[not] SBAR[whom we would deign to work for] 2) SBAR ->SBAR[when they are fine] ,[,] RB[only] SBAR[when they are mucked up or obscure]

Table 19: Top 10 production rules with examples for Sports

Positive Rules
$^{+}$ **WHNP->WP$ NN NN**
1) WHNP ->WP$[whose] NN[baseball] NN[career] 2) WHNP ->WP$[whose] NN[return] NN[date]
$^{+}$ **VP->VBD PRT , S**
1) VP ->VBD[left] PRT[off] ,[,] S[decisively winning the featured Copley Cup race of the 27th annual San Diego Crew Classic yesterday for the second consecutive year] 2) VP ->VBD[lashed] PRT[out] ,[,] S[accusing the league of racism]
$^{+}$ **NP->CD JJ JJ NN NN**
1) NP ->CD[one] JJ[infamous] JJ[dining] NN[hall] NN[brawl] 2) NP ->CD[seven] JJ[consecutive] JJ[first-round] NN[playoff] NN[series]
$^{+}$ **VP->ADVP VBD NP PP SBAR**
1) VP ->ADVP[out 95 seconds into the first round and Golota] VBD[left] NP[the arena] PP[in an ambulance] SBAR[after he lost consciousness in his locker room after the fight] 2) VP ->ADVP[quickly] VBD[switched] NP[him] PP[to second base] SBAR[because Chuck Knoblauch could not throw straight]
$^{+}$ **ADVP->JJ**
1) ADVP ->JJ[next] 2) ADVP ->JJ[free]

Negative Rules
$^{-}$ **NP->NP , CC NP , PP**
1) NP ->NP[Bob Brenly 's use] ,[,] CC[or] NP[overuse] ,[,] PP[of Curt Schilling] 2) NP ->NP[Vt.] ,[,] CC[minus] NP[a number of players still participating in the World Cup] ,[,] PP[including Gretzky , who has been Team Canada 's best player but is also the Rangers ' biggest question mark]
$^{-}$ **NP->DT MD CD NN**
1) NP ->DT[a] MD[May] CD[31] NN[deadline] 2) NP ->DT[a] MD[March] CD[4] NN[night]
$^{-}$ **NP->NP JJ NNP NN NN**
1) NP ->NP[Maryland 's] JJ[first] NNP[A.C.C.] NN[tournament] NN[championship] 2) NP ->NP[the year 's] JJ[first] NNP[Grand] NN[Slam] NN[tournament]
$^{-}$ **NP->PRP$**
1) NP ->PRP$[his] 2) NP ->PRP$[its]
$^{-}$ **XS->JJ IN**
1) XS ->JJ[much] IN[over] 2) XS ->JJ[further] IN[than]

Table 20: Top 10 production rules with examples for Politics

Positive Rules
$^{+}$ **NP->DT NNP : NNP NNP**
1) NP ->DT[the] NNP[Editor] :[:] NNP[Philip] NNP[Gourevitch] 2) NP ->DT[the] NNP[Editor] :[:] NNP[Henry] NNP[Siegman]
$^{+}$ **NP->DT VBG NNP NNP**
1) NP ->DT[the] VBG[collapsing] NNP[Soviet] NNP[Union] 2) NP ->DT[the] VBG[ruling] NNP[Communist] NNP[Party]
$^{+}$ **VP->VBG NP PRT SBAR**
1) VP ->VBG[propelling] NP[a civic debate] PRT[over] SBAR[whether to change the way Americans experience and ultimately build urban public spaces] 2) VP ->VBG[provoking] NP[a debate] PRT[about] SBAR[whether American courts would repeat the kinds of rulings that restricted the civil rights of Japanese-Americans during World War II]
$^{+}$ **VP->VP CC VP S**
1) VP ->VP[are being held in Banco Delta Asia in Macao] CC[and] VP[are] S[to be transferred to a North Korean account at the Bank of China] 2) VP ->VP[said the contacts were informal] CC[and] VP[had no bearing on the efforts] S[to help him settle in Panama]
$^{+}$ **ADVP->ADVP CC ADVP**
1) ADVP ->ADVP[at least another week] CC[and] ADVP[perhaps longer] 2) ADVP ->ADVP[far enough] CC[and] ADVP[well enough]

Negative Rules
$^{-}$ **ADJP->JJ CC RB JJ**
ADJP ->JJ[important] CC[but] RB[relatively] JJ[routine] ADJP ->JJ[tragic] CC[but] RB[not] JJ[surprising]
$^{-}$ **ADVP->DT RP**
ADVP ->DT[all] RP[over] ADVP ->DT[all] RP[around]
$^{-}$ **NP->NP NN PP S**
NP ->NP[Asmat Ali Janbaz 's] NN[explanation] PP[for the American military helicopters] S[flying over this isolated mountain valley last Thursday afternoon] NP ->NP[the Chinese Government 's] NN[use] PP[of military force] S[to suppress the 1989 Tiananmen demonstrations]
$^{-}$ **SBAR->WHADJP S**
SBAR ->WHADJP[exactly what] S[you were doing when you heard that Franklin D. Roosevelt had died , or that John F. Kennedy had been shot , or that Martin Luther King Jr. was dead] SBAR ->WHADJP[How delightful] S[it must be these days to be a member of the Chinese Communist Politburo]
$^{-}$ **NP->VBN NNP NNS**
NP ->VBN[suspected] NNP[Qaeda] NNS[members] NP ->VBN[suspected] NNP[Taliban] NNS[fighters]

References

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

Ashok, V. G., Feng, S., & Choi, Y. (2013). Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1753–1764.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*(1), pp. 32–68.

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter.. *ICWSM*, *11*, 438–441.

Berg-Kirkpatrick, T., Gillick, D., & Klein, D. (2011). Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 481–490. Association for Computational Linguistics.

Bertolami, R., & Bunke, H. (2006). Early feature stream integration versus decision level combination in a multiple classifier system for text line recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, Vol. 2, pp. 845–848.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Cook, P., & Hirst, G. (2013). Automatically assessing whether a text is clichéd, with applications to literary analysis. *Proceedings of NAACL HLT 2013.*

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., & Lee, L. (2009). How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of WWW*, pp. 141–150.

de Marneffe, M., Manning, C. D., & Potts, C. (2012). Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, *38*(2), 301–333.

Feng, X., Huang, L., Tang, D., Ji, H., Qin, B., & Liu, T. (2016). A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers.*

Ganjigunte Ashok, V., Feng, S., & Choi, Y. (2013). Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1753–1764.

Gillick, D., & Favre, B. (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pp. 10–18. Association for Computational Linguistics.

Gillick, D., Riedhammer, K., Favre, B., & Hakkani-Tur, D. (2009). A global optimization framework for meeting summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4769–4772. IEEE.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23–62.

Jurafsky, D., Ranganath, R., & McFarland, D. A. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, pp. 638–646.

Kahneman, D. (2011). *Thinking, Fast and Slow*.

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

Louis, A., & Nenkova, A. (2013). What makes writing great? first experiments on article quality prediction in the science journalism domain. *TACL*.

Louis, A., & Nenkova, A. (2014). Verbose, laconic or just right: A simple computational model of content appropriateness under length constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pp. 636–644.

Malmasi, S., & Dras, M. (2014). Chinese native language identification. In *Proceedings of EACL*, Vol. 2, pp. 95–99.

Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., & Sundheim, B. (2002). Summac: a text summarization evaluation. *Natural Language Engineering*, *8*(1), 43–68.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60.

Metallinou, A., Lee, S., & Narayanan, S. (2010). Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA*, pp. 2462–2465.

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

Nenkova, A., & Louis, A. (2008). Can you summarize this? identifying correlates of input difficulty for multi-document summarization. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pp. 825–833.

Nguyen, T. H., & Grishman, R. (2016). Modeling skip-grams for event detection with convolutional neural networks. In *EMNLP*, pp. 886–891. The Association for Computational Linguistics.

Over, P., Dang, H., & Harman, D. (2007). Duc in context. *Inf. Process. Manage.*, *43*(6), 1506–1520.

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*, 1–17.

Peng, H., Song, Y., & Roth, D. (2016). Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 392–402.

Post, M., & Bergsma, S. (2013). Explicit and implicit syntactic features for text classification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 866–872.

Queneau, R. (1947). *Exercises in Style*.

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

Raaijmakers, S., Truong, K., & Wilson, T. (2008). Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 466–474.

Saurí, R., & Pustejovsky, J. (2009). Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, *43*(3), 227–268.

Sinclair, J., & Ball, J. (1996). *Preliminary recommendations on text typology.*

Tulyakov, S., Jaeger, S., Govindaraju, V., & Doermann, D. (2008). Review of classifier combination methods. In *In Machine Learning in Document Analysis and Recognition. Informatica 34 (2010) 111?118 S. Vemulapalli et al.*

van Halteren, H., Zavrel, J., & Daelemans, W. (1998). Improving data driven wordclass tagging by system combination. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pp. 491–497.

Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science.*

Wilson, M. (1988). The mrc psycholinguistic database: Machine readable dictionary. *Behavioural Research Methods, Instruments and Computer, version 2*, *20*(1), 6–11.

Yang, Y., Bao, F., & Nenkova, A. (2017). Detecting (un)important content for single-document news summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 707–712.

Yang, Y., & Nenkova, A. (2014). Detecting information-dense texts in multiple news domains. In *Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence.*

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*, pp. 129–136.