Accurate Argumentative Zoning with Maximum Entropy models

Stephen Merity and **Tara Murphy** and **James R. Curran**
School of Information Technologies
University of Sydney
NSW 2006, Australia
{smerity,tm,james}@it.usyd.edu.au

Abstract

We present a maximum entropy classifier that significantly improves the accuracy of *Argumentative Zoning* in scientific literature. We examine the features used to achieve this result and experiment with Argumentative Zoning as a sequence tagging task, decoded with Viterbi using up to four previous classification decisions. The result is a 23% F-score increase on the Computational Linguistics conference papers marked up by Teufel (1999).

Finally, we demonstrate the performance of our system in different scientific domains by applying it to a corpus of Astronomy journal articles annotated using a modified Argumentative Zoning scheme.

1 Introduction

The task of generating automatic summarizations of one or more texts is a central problem in Natural Language Processing (NLP). Summarization is a fundamental component for future information retrieval and question answering systems, incorporating both natural language understanding and natural language generation.

Comprehension-based summarization, e.g. Kintsch and Van Dijk (1978) and Brown et al. (1983), is the most ambitious model of automatic summarization, requiring a complete understanding of the text. Due to the failure of rule-based NLP and knowledge representation, other less knowledge-intensive methods now dominate.

Sentence extraction, e.g. Brandow et al. (1995) and Kupiec et al. (1995), selects a small number of abstract worthy sentences from a larger text. The resulting sentences form a collection of excerpt sentences meant to capture the essence of the text. The next stage is information fusion (Barzilay et al., 1999; Knight and Marcu, 2000) which attempts to combine the excerpts into a more cohesive text. These methods can create inflexible and incoherent extracts that result in under-informative results (Teufel et al., 1999).

Argumentative Zoning (Teufel, 1999; Teufel and Moens, 2002) attempts to solve this problem by representing the structure of a text using a rhetorically-based schema. Sentences are classified into one of a small number of non-hierarchical argumentative roles, which can then be used in both the sentence extraction and text generation/fusion phase of automatic summarization. Argumentative Zoning can enable tailored summarizations depending on the needs of the user, e.g. a layperson versus a domain expert.

The first experiments in Argumentative Zoning used Naïve Bayes (NB) classifiers (Kupiec et al., 1995; Teufel, 1999) which assume conditional independence of the features. However, this assumption is rarely true for the kinds of rich feature representations we want to use for most NLP tasks.

Maximum entropy (ME) models have become popular in NLP because they can incorporate evidence from the complex, diverse and overlapping features needed to represent language. Some example applications include part-of-speech (POS) tagging (Ratnaparkhi, 1996), parsing (Johnson et al., 1999), language modelling (Rosenfeld, 1996), and text categorisation (Nigam et al., 1999).

We have developed an Argumentative Zoning (*zone*) classifier using a ME model. We compare our zone classifier to a reimplementation of Teufel and Moens (2002)'s NB classifier and features on their original Computational Linguistics corpus. Like Teufel (1999), we model zone classification as a sequence tagging task. Our zone classifier achieves an F-score of 96.88%, a 20% improvement. We also show how Argumentative Zoning can be applied to other domains by evaluating our system on a corpus of Astronomy journal articles, achieving an F-measure of 97.9%.

Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009, pages 19–26,
Suntec, Singapore, 7 August 2009.

Category	Abbr.	Description
Background	**BKG**	general scientific background
Other	**OTH**	neutral descriptions of other researcher's work
Own	**OWN**	neutral descriptions of the authors' new work
Aim	**AIM**	statements of the particular aim of the current paper
Textual	**TXT**	statements of textual organisation of the current paper
Contrast	**CTR**	contrastive or comparative statements about other work explicit mention of weaknesses of other work
Basis	**BAS**	statements that own work is based on other work

Table 1: Teufel's (1999) Argumentative Zones

2 Argumentative Zoning

Teufel (1999) introduced a new rhetorical analysis for scientific texts called *Argumentative Zoning*. Each sentence of an article from the scientific literature is classified into one of seven basic rhetorical structures shown in Table 1.

The first three: Background, Other, and Own, are part of the basic schema and represent attribution of intellectual ownership. The four additional categories: aim, textual, contrast, and basis, are based upon Swales (1990)'s Creating A Research Space (CARS) model, and provide pointed information about the author's stance and the paper itself. Teufel assumes that each sentence only requires a single classification and that all sentences clearly fit into the above structure. The assumption is clearly not always correct, but is a useful approximation nevertheless.

Due to the specific nature of these classifications it is hoped that this will allow for much more robust automatic abstraction generation. Summaries of a paper could be created specifically for the user, either focusing on the aim of the work, the work's stance in the field (what other works it is based upon or compared with) and so on.

Teufel used Argumentative Zoning to determine the author's use and opinion of other authors they cite in their work and also to create *Rhetorical Document Profiles* (RDP), a type of summarization used to provide typical information that a new reader may need in a systematic manner.

For the use of Argumentative Zoning in RDPs Teufel (1999) points out that due to the redundancy in language that near perfect accuracy is not required as important pieces of information will be repeated in the paper. Recognising these salient points once is enough for them to be included in the RDP. In further tasks, such as the analysis of the function of citations (Teufel et al., 2006) and automatic summarization, higher levels of accuracy are more critical.

3 Maximum Entropy models

Maximum entropy (ME) or log-linear models are statistical models that can incorporate evidence from a diverse range of complex and potentially overlapping features. Unlike Naïve Bayes (NB), the features can be conditionally dependent given the class, which is important since feature sets in NLP rarely satisfy this independence constraint.

The ME classifier uses models of the form:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(x,y)\right) \qquad (1)$$

where y is the zone label, x is the context (the sentence) and the $f_i(x,y)$ are the *features* with associated weights λ_i.

The probability of a sequence of zone labels $y_1 \ldots y_n$ given a sequence of sentences is $s_1 \ldots s_n$ is approximated as follows:

$$p(y_1 \ldots y_n | s_1 \ldots s_n) \approx \prod_{i=1}^{n} p(y_i|x_i) \qquad (2)$$

where x_i is the context for sentence s_i. In our experiments that treat argumentative zoning as a sequence labelling task, the context x_i incorporates history information – i.e. the previous labelling decisions of the classifier. Optimal decoding of this sequence uses the Viterbi algorithm, which we compare against the Oracle case of knowing the correct label for the previous sentence.

The features are binary valued functions which pair a zone label with various elements of the sentential context; for example:

$$f_j(x,y) = \begin{cases} 1 \text{ if } \texttt{goal} \in x \ \&\ y = \texttt{AIM} \\ 0 \text{ otherwise} \end{cases} \qquad (3)$$

$\texttt{goal} \in x$, that is, the word `goal` is part of the context of the sentence, is a *contextual predicate*.

The central idea in maximum entropy modelling is that the model chosen should satisfy all of the constraints imposed by the training data (in the

form of empirical feature counts from the training data) whilst remaining as unbiased as possible. This is achieved by selecting the model with the maximum entropy, i.e. the most uniform distribution, given the constraints.

Our classifier uses the maximum entropy implementation described in Curran and Clark (2003). Generalised Iterative Scaling (GIS) is used to estimate the values of the weights and we use a Gaussian prior over the weights (Chen and Rosenfeld, 1999) which allows many rare, but informative, features to be used without overfitting. This will be an important property when we use sparse features like bigrams in the models below.

4 Modelling Argumentative Zones

4.1 Our Features

The two primary sources of features for our zone classifier were the words in the sentences and the position of the sentence relative to the rest of the paper. A number of feature types use additional external resources (e.g. semantic lists of agents or common rhetorical patterns) or annotations (e.g. named entities). Where feasible we have reimplemented the features described in Teufel (1999). In other cases, our features are somewhat simpler.

Since the Curran and Clark (2003) classifier only accepts binary features, any numerical features had to be bucketed into smaller sets of alternatives to reduce sparseness, either by integer division or through reducing the number by scaling to a small integer range. The features we implemented are described below.

Unigrams, bigrams and n-grams

A sub-sequence of *n* words from a given sentence. We include unigram and bigram features and report them individually and together (as n-grams). These features include all of the unigrams and bigrams above the feature cutoff, unlike Teufel's cont-1 features below. Also, both the Computational Linguistics and Astronomy corpora contain marked up citations, cross-references to tables, figures, and sections and mathematical expressions. In the Computational Linguistics corpus self citations are distinguished from other citations. These structured elements have been normalised to a single token each, e.g. __CITE__. These tokens have been retained in the unigram and bigram features.

first The first four words of a sentence, added individually.

Sections, positions, and lengths

section A section counter which increments on each heading to measure the distance into the document. It does not take into consideration whether they are sub-headings or similar. There are two versions of this feature. The first is a straight counter (1 to n) and the second is grouped into two buckets representing each half of the paper (breaking at the middle section).

location The position of a sentence between two headings (representing a section). There are two versions of this feature, one counts to a maximum of 10 and the other represents a percentage through the section bucketed into 20% intervals.

paragraph The position of the sentence within a paragraph. Again there are two features – either straight counts (with a maximum of 10) or bucketed into thirds of a paragraph.

length of sentence grouped into multiples of 3.

Named entity features

Our astronomy corpus has been manually annotated with domain-specific named entity information (Murphy et al., 2006). There are 12 coarse-grained categories and 43 fine-grained categories including star, galaxy, telescope, as well as a number of the usual categories including person, organisation and location. Both the coarse-grained and fine-grained categories were used as features.

4.2 Teufel (1999)'s features

To compare with previous work, we also implemented most of the features that gave Teufel (1999) the best performance. We list all of the feature types in Table 2, indicating which ones have and have not been implemented.

Teufel's unigram features (**cont-1**) are filtered using TF-IDF to select the top scoring 10 words in each document, and then these are used to mark the top 40 sentences in each document containing those filtered words.

TLoc marks the position of the sentence over the entire paper, using 10 unevenly sized segments (larger segments are in the middle of the paper).

Struct-1 marks where a sentence appears in a section. It divides each section into three equally sized segments; singles out the first and the last sentence as separate segments; the second and

Name	Impl?	Description
Cont-1	yes	An application of TF-IDF over the words and sentences
Cont-2	partial	Does the sentence contain words in the title or heading (excluding stop words)
TLoc	yes	Position of the sentence in relation to 10 segments (A-J)
Struct-1	yes	Position within a section
Struct-2	yes	Relative position of sentence within a paragraph
Struct-3	partial	Type of headline of the current section
TLength	yes	Is the sentence longer than 15 words?
Syn-1	no	Voice of the first finite verb in the sentence
Syn-2	no	Tense of the first finite verb in the sentence
Syn-3	no	Is the first finite verb modified by a modal auxiliary
Cit-1	yes	Does the sentence contain a citation or name of author?
Formu	yes	Does a formulaic expression occur in the sentence
Ag-1	yes	Type of agent
Ag-2	yes	Type of action (with or without negation)

Table 2: Teufel (1999)'s set of features

third sentence as a sixth segment; and the second-last plus third-last sentence as a seventh segment. **Struct-3** the type of section heading for the current section. In our case, we have not mapped these down to the reduced set used by Teufel.

Formu uses pattern matching rules to identify formulaic expressions. **Ag-1** and **Ag-2** identify agent and action expressions from gazetteers. Teufel (1999) provides these in the appendices.

4.3 Feature Cutoff

Features that occur rarely in the training set are problematic because the statistics extracted for these features are not reliable. They may still contribute positively to the ME model because we use Gaussian smoothing (Chen and Rosenfeld, 1999) help avoid overfitting.

Instead of including every possible feature, we used a cutoff to remove features that occur less than four times. This primarily applies to the n-gram features, especially bigrams, which were quite sparse given the small quantity of training data. Due to the speed of the ME implementation it is possible to have quite a low cut-off.

4.4 History features and Viterbi

In order to take advantage of the predictability of tags given prior sequences (for example, AIM commonly following itself) we used history features and treated Argumentative Zoning as a sequence labelling task. Since each prediction now relies on the previous decisions we used the Viterbi algorithm to find the optimal sequence.

Given the small number of labelling alternatives, we experimented with several history lengths ranging from previous label to the previous four labels. To determine the impact of this feature in an ideal situation, we also experimented with using an Oracle set of history features.

5 Results

Our results are produced using ten-fold cross validation and are reported in terms of precision, recall and f-score for each of the zone classes, and a weighted average over all classes. We have investigated the impact of each feature type using subtractive analysis, where we have also calculated paired t-test confidence intervals (the error values reported are the 95% confidence interval).

The baselines for both sets were already quite high (at least 70%) due to the common tag of OWN, representing the author's own work, but our results show significant improvements over this baseline.

5.1 CMP-LG Corpus

The CMP-LG corpus is a collection of 80 conference papers collected by Teufel (1999) from the Computation and Language E-Print Archive [1]. The LaTeX source was converted to HTML with Latex2HTML then transformed into XML with custom PERL scripts. This text was then tokenized using the TTT (Text Tokenization) System into Penn Treebank format. The result is a corpus of 12,000 annotated sentences, containing 333,000 word tokens, in XML format.

We attempted to recreate Teufel's original experiments by emulating the features she used with the same type of classifier. We used Weka's (Frank et al., 2005) implementation of the NB classifier.

Table 3 reproduces the results from Teufel and Moens (2002) alongside our reimplementation of

[1] http://xxx.lanl.gov/cmp-lg/

Tag	original			reproduced		
	P	R	F	P	R	F
AIM	44	65	52	45.8	57.8	51.1
BAS	37	40	38	23.8	37.0	28.9
CTR	34	20	26	33.1	19.2	24.3
BKG	40	50	45	46.9	53.6	50.1
OTH	52	39	44	70.6	55.0	61.8
TXT	57	66	61	66.3	47.6	55.4
OWN	84	88	86	86.7	90.8	88.7
Weighted	72	73	72	76.8	76.8	76.8

Table 3: Teufel and Moens (2002)'s and our NB performance on CMP-LG

History Type	Order	Performance	
Baseline	None	93.16	
Viterbi	First	1.77	± 0.49%
Viterbi	Second	1.97	± 0.42%
Viterbi	Third	2.08	± 0.45%
Viterbi	Fourth	2.1	± 0.46%
Viterbi	Fifth	2.13	± 0.46%
Oracle	First	3.67	± 0.68%
Oracle	Second	4.06	± 0.70%

Table 4: History features on the CMP-LG corpus with ME model of unigram/bigram features only

Feature	Classifier	Viterbi
Ngrams	-21.39±2.35%	-23.23±3.24%
Unigram	-8.00±1.02%	-7.53±1.14%
Bigram	-7.89±1.20%	-6.87±1.44%
Concept	-0.06±0.24%	-0.06±0.16%
First	-1.24±0.44%	-1.14±0.39%
Length	-0.34±0.24%	-0.40±0.25%
Section	-0.42±0.27%	-0.27±0.33%
Location	0.03±0.20%	0.04±0.07%
Paragraph	0.10±0.15%	0.01±0.08%
All	95.69%	96.88%

Table 5: Subtractive analysis CMP-LG ME model

the features using Weka's NB classifier. We have been able to replicate their results to a reasonable extent – gaining higher overall performance using most of their original features. Notably, our Other class is significantly more accurate whilst the original Basis class did better.

Our next experiment investigated the value of treating Argumentative Zoning as a sequence labelling task, i.e. the impact of the Markov history features and Viterbi decoding on performance. For these experiments we only used the unigram and bigram features with the maximum entropy classifier. Table 4 presents the results: the baseline is already much higher than the NB classifier which is a result of both the unigram/bigram features and the ME classifier itself.

The improvement using longer Markov windows (up to 2.13%) is also shown – and longer windows are better, although there is diminishing returns. We chose a Markov history of the four previous decisions for the rest of our experiments. Table 4 also shows that knowing the previous label perfectly (with the Oracle experiment) can make a large difference to classification accuracy.

Feature	Change
TLength	-2.09±9.96%
Struct-1	0.38±6.08%
TLoc	0.96±7.25%
Struct-3	-1.65±6.76%
Cont-2	-1.10±6.39%
Struct-2	1.59±7.99%
Ag-1/2	-0.39±8.97%
Formu	0.14±8.46%
Cit-1	-1.88±5.19%
Cont-1	-0.38±5.85%
All	70.25%

Table 6: Teufel's Subtractive analysis CMP-LG ME

Table 5 presents the subtractive analysis to determine the impact of different feature types. From this we can see that the n-grams (unigrams and bigrams) have by far the largest impact – and neither of these feature types was directly implemented by Teufel and Moens (2002). The next most important features are the first few words (again a unigram type feature), length and the section number. The Markov history features also have an impact of just over 1%.

Table 6 shows a different story for Teufel's features using the maximum entropy model. It seems that none of the feature types alone are making an enormous contribution and that the impact of them varies enormously between folds (the confidence intervals are far bigger than the differences).

Finally, Table 7 gives the results of using the maximum entropy model with Markov history length four and all of the features. Overall, we improve Teufel and Moens' performance by just under 20% on our reproduced experiments.

5.2 Astronomical Corpus

The astronomical corpus was created by Murphy et al. (2006) and consists of papers obtained from arXiv (2005)'s astrophysics section (astro-ph). The papers were converted from LaTeX to Unicode by a custom script which attempted to retain as much of the paper's special characters and formatting as possible.

The resulting text was then processed using MXTerminator (Reynar and Ratnaparkhi, 1997) with an additional Python script to find sentence

Category	Abbr.	Description
Background	**BKG**	As has been noted in prior studies , Abell|GXYC 2255|GXYC has an unusually large number of galaxies with extended radio emission .
Other	**OTH**	This is consistent with the findings of Hogg|P Fruchter|P (1999|DAT) who found that GRB hosts are in general subluminous galaxies .
Own	**OWN**	We scanned the data of about 1.8|DUR year|DUR (TJDs|DUR 11000-11699|DUR) and found 30 new GRB-like events .
Data	**DAT**	In Fig . __REF__ we present the 1.4|FRQ GHz|FRQ radio images of the cluster A2744|GXYC , at different angular resolutions . *(subclassed from OWN)*
Observation	**OBS**	Smith|P et al. (2001|DAT) reported no detection of transient emission at sub-mm (850|WAV um|WAV) wavelengths . *(subclassed from OTH)*
Technique	**TEC**	Reduction of the NIR images was performed with the IRAF|CODE and STSDAS|CODE packages . *(subclassed from OWN)*

Figure 1: Examples of sentences with the given tags in the astronomical corpus

Tag	P	R	F
AIM	96.5	88.2	92.2
BAS	86.7	89.8	88.2
CTR	92.1	89.0	90.5
BKG	86.0	96.3	90.9
OTH	96.3	91.7	93.9
TXT	98.2	93.8	95.9
OWN	98.6	99.2	98.9
Weighted	96.88	96.88	96.88

Table 7: Final CMP-LG ME performance

Feature	Classifier	Viterbi
Ngrams	-18.83±3.74%	-16.03±2.99%
Unigram	-5.51±1.37%	-5.25±2.00%
Bigram	-2.04±0.78%	-1.79±0.87%
Concept	-0.18±0.29%	-0.05±0.12%
Entity	-0.18±0.39%	-0.31±0.23%
First	-0.02±0.29%	-0.86±0.79%
Length	-0.06±0.16%	-0.08±0.10%
Paragraph	-0.04±0.20%	0.07±0.19%
Section	-0.29±0.24%	-0.40±0.57%
Location	-0.09±0.25%	0.06±0.15%
All	98.15%	96.68%

Table 8: Subtractive analysis ASTRO ME model

boundaries, and then tokenized using the Penn Treebank (Marcus et al., 1993) sed script, with another Python script fixing common errors. The LaTeX, which the tokenizer split off incorrectly, was then reattached.

Each sentence of the corpus was then annotated using a modified version of the Argumentative Zoning schema. While the original three zones: Background, Own, Other are used, we have replaced the CARS labels with content labels describing aspects of the work: **DAT** for data used in the analysis, **OBS** for observations performed, and **TEC** for techniques applied. Only Own and Other are subclassed with the extended schema of Data, Observation and Techniques. Examples of each zone classification are shown in Figure 1.

Tag	P	R	F
BKG	92.1	97.1	94.5
OTH	95.0	97.1	96.1
OTH-DAT	100.0	92.3	96.0
OTH-OBS	91.3	93.3	92.3
OTH-TEC	100.0	100.0	100.0
OWN	99.9	99.3	99.6
OWN-DAT	95.9	86.6	91.0
OWN-OBS	98.2	89.4	93.6
OWN-TEC	90.4	100.0	94.9
Weighted	97.9	97.9	97.9

Table 9: Final ASTRO ME model performance

Table 8 shows the impact of different feature types on classification accuracy for the Astronomy corpus. Again, the most important features are the n-grams (although to a slightly lesser extent than for the Computational Linguistics corpus). The other features make very little contribution at all. Disappointingly, the (gold-standard) named entity features contribute very little additional information – which is surprising given that the content categories (data and observation) are directly connected with some of the entity types (like telescope).

In the Astronomy corpus, the Markov history features actually have a detrimental effect, which suggests the history is misleading. This warrants further exploration, but we suspect there may be more changing backwards and forwards between argumentative zones in the Astronomy corpus. Overall, we can see that the two tasks are of a similar level of difficulty of around 96% F-score.

Table 9 shows the distribution over zones and content labels for the Astronomy corpus. The Background label is the hardest to reproduce even though it is not split into content sub-types. The sub-types are relatively rare for Other, so the results should not be considered as reliable.

Tag	P	R	F
BKG NB CMP-LG	51.5%	61.1%	55.9%
OTH NB CMP-LG	73.0%	64.2%	68.3%
OWN NB CMP-LG	91.9%	93.1%	92.5%
BKG NB ASTRO	63.1%	63.5%	63.3%
OTH NB ASTRO	53.9%	39.7%	45.7%
OWN NB ASTRO	88.5%	93.0%	90.7%
BKG ME CMP-LG	53.6%	27.5%	36.3%
OTH ME CMP-LG	63.0%	24.4%	35.2%
OWN ME CMP-LG	81.7%	96.8%	88.6%
BKG ME ASTRO	61.2%	29.5%	39.8%
OTH ME ASTRO	50.4%	20.0%	28.6%
OWN ME ASTRO	81.2%	96.7%	88.2%

Table 10: Comparing CMP-LG and ASTRO directly on the basic annotation scheme

Table 10 compares the performance of our Naïve Bayes and Maximum Entropy classifiers on the two corpora for just the basic annotation scheme: Background, Own and Other. The features used are the set of Teufel features we have implemented (so it does not include unigram or bigram features).

The results show that classifiers for both corpora behave in quite similar ways on the basic scheme. Own is by far the most frequent category, and not surprisingly, it is most accurately classified in both domains. Background seems to be easier to distinguish in Astronomy, but Other is more distinct in Computational Linguistics.

Further, we see no advantage to using maximum entropy models over Naïve Bayes when the feature set is not sophisticated/overlapping enough, and the dataset large enough, to warrant the extra power (and cost).

6 Conclusion

This paper has presented new models of Argumentative Zoning using Maximum Entropy (ME) models. We have demonstrated that using ME models with standard word features, such as unigrams and bigrams, significantly outperforms Naïve Bayes models incorporating task-specific features. Further, these task-specific features had very little additional impact on the ME model.

Our ME model has raised the state-of-the-art in automatic Argumentative Zoning classification from 76% to 96.88% F-score on Teufel's Computational Linguistics conference paper corpus.

To test the wider applicability of Argumentative Zoning, we have annotated a corpus of Astronomy journal articles with a modified zone and content scheme, and achieved a similar level of performance using our maximum entropy classifier. We found that more sophisticated semantic features, e.g. gold-standard named entities, also had little impact on the accuracy of our classifier.

Now that we have a very accurate Argumentative Zone classifier, we would like to investigate the impact of Argumentative Zones in information retrieval, question answering, and summarization tasks, particularly in the astronomy domain, where we have additional tools such as the named entity recognizer.

In summary, using a maximum entropy classifier with simple unigram and bigram features results in a very accurate classifier for Argumentative Zones across multiple domains.

Acknowledgements

We would like to thank Sophie Liang and the anonymous reviewers for their helpful feedback on this paper. This work has been supported by the Australian Research Council under Discovery project DP0665973. The first author was supported by the Microsoft Research Asia Scholarship in IT at the University of Sydney.

References

arXiv. 2005. arxiv.org archive. `http://arxiv.org`.

R. Barzilay, K.R. McKeown, and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 550–557. Association for Computational Linguistics Morristown, NJ, USA.

R. Brandow, K. Mitze, and L.F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and management*, 31(5):675–685.

A.L. Brown, J.D. Day, and R.S. Jones. 1983. The development of plans for summarizing texts. *Child Development*, pages 968–979.

Stanley Chen and Ronald Rosenfeld. 1999. A Gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University, Pittsburgh, PA.

James R. Curran and Stephen Clark. 2003. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 91–98, Budapest, Hungary, 12–17 April.

E. Frank, M.A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I.H. Witten, and L. Trigg. 2005. Weka-a machine learning workbench for data mining. *The Data Mining and Knowledge Discovery Handbook*, pages 1305–1314.

M. Johnson, S. Geman, S. Canon, Z. Chi, and S. Riezler. 1999. Estimators for stochastic 'unification-based' grammars. In *Proceedings of the 37th Meeting of the ACL*, pages 535–541, University of Maryland, MD.

W. Kintsch and T.A. Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363–94.

K. Knight and D. Marcu. 2000. Statistics-based summarization-step one: Sentence compression. In *Proceedings of the National Conference on Artificial Intelligence*, pages 703–710. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

J. Kupiec, J. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM New York, NY, USA.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

T. Murphy, T. McIntosh, and J.R. Curran. 2006. Named entity recognition for astronomy literature. In *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW)*.

K. Nigam, J. Lafferty, and A. McCallum. 1999. Using maximum entropy for text classification. In *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, Stockholm, Sweden.

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the EMNLP Conference*, pages 133–142, Philadelphia, PA.

J.C. Reynar and A. Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the fifth conference on Applied natural language processing*, pages 16–19.

R. Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228.

J.M. Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

S. Teufel and M. Moens. 2002. Summarising scientific articles — experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

S. Teufel, J. Carletta, and M. Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL 1999*.

S. Teufel, A. Siddharthan, and D. Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110.

S. Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh, Edinburgh, UK.