Document and Corpus Level Inference For Unsupervised and Transductive Learning of Information Structure of Scientific Documents

Roi Reichart   Anna Korhonen
The Computer Laboratory, University of Cambridge, UK
Roi.Reichart@cl.cam.ac.uk, alk23@cam.ac.uk

ABSTRACT
Inferring the information structure of scientific documents has proved useful for supporting information access across scientific disciplines. Current approaches are largely supervised and expensive to port to new disciplines. We investigate primarily unsupervised discovery of information structure. We introduce a novel graphical model that can consider different types of prior knowledge about the task: within-document discourse patterns, cross-document sentence similarity information based on linguistic features, and prior knowledge about the correct classification of some of the input sentences when this information is available. We apply the model to Argumentative Zoning (AZ) scheme and evaluate it on a fully unsupervised learning scenario and two transduction scenarios where the categories of some test sentences are known. The model substantially outperforms similarity and topic model based clustering approaches as well as traditional transduction algorithms.

TITLE AND ABSTRACT IN FINNISH

Dokumentti- ja korpustason inferenssiin perustuva ohjaamattomankoneoppimisen tekniikka tieteellisen julkaisujen rakenteen analyysissa

Tieteellisten julkaisujen rakenteen analyysi voi tukea tietojen saatavuutta eri tieteenaloilta. Nykyiset koneoppimismetodit ovat pitkälti ohjattuja ja niiden soveltaminen uusille tieteenaloille on kallista. Tämä artikkeli tutkii pääasiassa ohjaamatonta julkaisujen rakenteen analyysia. Lähtökohtana on uusi graafinen malli, joka pystyy integoimaan erilaista etukäteistietoa tehtävästä: dokumenttien sisäisen diskurssin, dokumenttienvälisten samankaltaisuuden kielellisten ominaisuuksien suhteen, ja tietoa joidenkin lauseiden oikeasta luokittelusta, silloin kun tämänkaltaista tietoa on saatavilla. Malli sovellettiin Argumentative Zoning (AZ) -analyysiin ja sen soveltuvuutta täysin ohjaamattomaan oppimiseen sekä transduktio-oppimiseen, jossa joidenkin testilauseiden luokat on tiedossa, tutkittiin. Malli osoittautuu huomattavasti tarkemmaksi kuin samankaltaisuuteen ja klusterointiin perustuvat vertailumallit sekä perinteiset transduktio-algoritmit.

KEYWORDS: Information structure, Argumentative Zoning , Approximate Inference.

FINNISH KEYWORDS: Rakenteen analyysi, Argumentative Zoning , Approksimoitu Inferenssi.

Proceedings of COLING 2012: Posters, pages 995–1006,
COLING 2012, Mumbai, December 2012.

1 Introduction

Information structure of scientific literature (i.e. the way scientists communicate their ideas, methods, results, conclusions, and so forth, to their audience) has been a topic of intense research within different disciplines (Taboada and Mann, 2006; Argamon et al., 2008; Deane et al., 2008; Lungen et al., 2010). Within Natural Language Processing (NLP), various schemes have been proposed for describing the information structure of scientific documents. These have been based on, for example, section names found in documents (Lin et al., 2006; Hirohata et al., 2008), rhetorical or argumentative zones (AZ) of sentences (Teufel and Moens, 2002; Mizuta et al., 2006; Teufel et al., 2009), qualitative aspects of scientific information (Shatkay et al., 2008) or core scientific concepts (Liakata et al., 2010).

Previous works have shown that it is possible to classify sentences in scientific documents according to the categories of such schemes (e.g. the Background, Method, Results and Conclusions categories of the AZ scheme) using supervised methods. These methods perform very well and their output has proved useful for important tasks such as information retrieval and extraction (Teufel, 2001; Teufel and Moens, 2002; Mizuta et al., 2006; Tbahriti et al., 2006; Ruch et al., 2007). This comes, however, with a heavy cost of requiring thousands of manually annotated sentences to achieve good performance. Even the weakly supervised approach by (Guo and Korhonen, 2011) requires hundreds of annotated sentences for optimal performance.

In this paper we focus on a primarily unsupervised approach to inferring information structure which avoids the high annotation cost of the supervised approaches. The only previous work on this topic that we are aware of is that of (Varge et al., 2012) who proposed a simple word-level Latent Dirichlet Allocation (LDA) model to the task, assuming that the phenomenon is mostly lexical. As we show in this paper, the information structure of scientific documents is governed by a number of additional factors, which calls for a more expressive model.

We propose a more sophisticated and flexible model capable of integrating different types of task knowledge, depending on the knowledge available in a real-life situation. We investigate two scenarios: (1) the fully unsupervised scenario where no manually annotated sentences are available; and (2) the transductive scenario (Gammerman et al., 1998) where the classes of some of the test set sentences are given. The transductive scenario is of particular interest when some test time knowledge about the document collection is available that could benefit learning. Examples of such knowledge are lexical cues (e.g. key words associated with a database index) for test sentences from a particular target category or sentence annotations that can be obtained fast for a small fraction of test data (e.g. using mechanical turk annotators).

Our model can take into account three types of knowledge about the task: (1) within-document discourse patterns; (2) linguistic feature representation used to model cross-document sentence similarity; and (3) in the transductive scenario, prior knowledge about the correct classification of some of the input sentences. Importantly, none of these knowledge types are actually required by the model, but the flexibility of the model enables us to consider all of them or only a subset.

We formulate our approach as a graphical model that encodes sentence-level knowledge via single-vertex potentials and knowledge about sets of sentences, both within and between documents, via global potentials. While these potentials encode important linguistic properties, they complicate the inference process. We therefore apply a linear-programming (LP) relaxation method (Sontag et al., 2008) which approximates the maximum aposteriori (MAP) assignment of our model. In our experiments the algorithm provably finds the exact MAP assignment.

We compare the predictions of our model to those of argumentative zoning (AZ) – a widely used information structure scheme (Teufel and Moens, 2002) where the core categories are argued to be domain-independent and which has been used to analyse texts in various disciplines such as computational linguistics (Teufel and Moens, 2002), law (Hachey and Grover, 2006), biology (Mizuta et al., 2006) and chemistry (Teufel et al., 2009). We experiment with the only publicly available AZ corpus: the corpus of 792 biomedical abstracts by (Guo et al., 2010) which provides AZ annotations for 7886 sentences. Our experimental evaluation shows that the model outperforms traditional algorithms for both the unsupervised and the transductive setups. by a large margin Our results show that it is possible to infer high quality knowledge about the information structure of scientific documents even when only little or no human annotation effort is involved.

2 Previous Work

Machine Learning for Information Structure Nearly all previous work on automatic detection of information structure has relied on supervised algorithms and, consequently, on corpora consisting of thousands of manually annotated sentences (Teufel and Moens, 2002; Lin et al., 2006; Hirohata et al., 2008; Shatkay et al., 2008; Teufel et al., 2009; Guo and Korhonen, 2011).

Recently, (Varge et al., 2012) were the first to apply unsupervised learning to the information structure of scientific documents. They applied standard word-level LDA models to the IMRAD scheme for the biomedical domain (along with their own information structure scheme for the aerospace domain). This purely lexical approach ignores other important linguistic phenomena, such as discourse patterns and syntactic properties, which play a role in information structure. The 35 F-score performance of their model indeed show that there is much scope for improvement. Our model integrates a much wider range of linguistic knowledge about the task at both within-document (e.g. discourse patterns) and cross-document (e.g. sentence similarity) levels, and can be flexibly applied to both fully unsupervised and transductive learning scenarios, depending on how much prior knowledge about the task is actually available. Although the transductive learning scenario can realistically occur when developing new corpora or applications, it has not been addressed in previous work on our task.

Corpus level Inference A number of recent models have obtained improved performance by sharing information between sentences and documents in large text collections (Sutton and McCallum, 2004; Taskar et al., 2002; Bunescu and Mooney, 2004; Finkel et al., 2005; Gupta et al., 2010; Rush et al., 2012; Reichart and Barzilay, 2012; Ganchev et al., 2010; Gillenwater et al., 2010; Mann and McCallum, 2010; Liang et al., 2009; Roth and Yih, 2005). We follow these works and model inter-sentence similarity across multiple scientific documents. To the best of our knowledge, this is the first model for the AZ classification task that explicitly shares information among sentences in different documents.

3 Model

Given a set of scientific documents our goal is to assign each sentence in these documents into a category that represents its role in the information structure of the document. As our data is biomedical, we use the version of the AZ scheme adapted for biology by (Mizuta et al., 2006). This version has ten zone categories. We focus on the five that appear in abstracts (as opposed to full papers): BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS.[1] For a detailed

[1] Two additional categories – RELATED and FUTURE work – appear only occasionally in abstracts. Since in our corpus only 2% of the sentences were tagged with one these categories we left their exploration for future work.

definition of the zone categories and an example annotated abstract see (Guo et al., 2010).

3.1 Model Structure

We denote the number of sentences in the document collection with n and the number of target categories with K. We define an undirected graphical model (Markov Random Field, MRF) with the vertex set $V = X \cup A$, where $X = \{x_1, \ldots, x_n\}$ consists of one vertex for every sentence in the document collection, and $A = \{a_1 \ldots a_K\}$ is a set of agreement vertices.

We integrate knowledge in the model through *singleton potentials* (defined over individual vertices) as well as *pairwise potentials* (defined between pairs of vertices). We consider the following types of knowledge: **(1) Within-Document Discourse Patterns** which encode the information conveyed by discourse patterns about the progress of information categories along a document. The discourse knowledge is encoded through within-document pairwise potentials as well as singleton potentials, both defined over vertices in X. **(2) Cross-Document Sentence Similarity** which encourages similar sentences in different documents to be assigned in the same category. This knowledge is encoded through cross-document pairwise potentials, defined over vertices in X, and through potentials between sentence vertices (X) and the agreement vertices (A). **(3) Class-Specific Lexical Cues** Encode lexical cues for the cluster of a given sentence through within-document pairwise potentials. **(4) Prior Knowledge on Sentence Categorization** which encodes prior knowledge about the categories of a predefined set of sentences through local potentials.

We use five types of potentials: (1) The singleton potentials encode context-free knowledge that does not depend on neighbouring vertices. (2) The pairwise potentials between sentence vertices in the same document encode discourse patterns that govern the information flow in the document. (3) The pairwise potentials between sentence vertices (X) in different documents encode the similarity between the sentences. The more similar a pair of sentences, the stronger the tendency of its members to be assigned to the same category. (4) The pairwise potentials between sentence vertices (X) and agreement vertices (A) encoded a tendency of sentences in different documents to be assigned to the same category based on sentence similarity patterns in their documents. In the last two potential types, the similarity between sentences is based on linguistic features. Finally, (5) the pairwise potentials between agreement vertices ensure that $category(a_i) = category(a_{i-1}) + 1$ by giving an infinite bonus to those assignments.

The resulting maximum aposteriori problem (MAP) takes the form of:

$$MAP(V) = \sum_{i=1}^{n} \theta_i(x_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} \theta_{i,j}(x_i, x_j) + \sum_{i=1}^{n}\sum_{j=1}^{K} \phi_{i,j}(x_i, a_j) + \sum_{i=1}^{K}\sum_{j=1}^{K} \xi_{i,j}(a_i, a_j)$$

We define the singleton and the pairwise potentials to take the following forms[2]:

$$\theta_i(x_i) = \left\{ \begin{array}{ll} \alpha & \text{if discourse pattern holds} \\ \infty & \text{if prior sentence-classification condition holds} \\ 0 & \text{otherwise} \end{array} \right\}$$

$$\theta_{i,j}(x_i, x_j) = \left\{ \begin{array}{ll} \beta & \text{if discourse pattern holds} \\ SimScore_{i,j} & \text{if similarity condition holds} \\ 0 & \text{otherwise} \end{array} \right\}$$

$$\phi_{i,j}(x_i, a_j) = \left\{ \begin{array}{ll} \gamma & \text{if similarity pattern holds} \\ 0 & \text{otherwise} \end{array} \right\}$$

Where $SimScore_{i,j}$ are the feature-based similarity scores computed between sentences in different documents and α, β and γ are the model parameters representing the relative strength

[2] To avoid clutter we omit the explicit definition of the pairwise potential between agreement nodes (ξ).

of the different types of knowledge. In section 3.2 we give a detailed description of the information encoded into these potentials.

3.2 Potentials and Encoded Knowledge

Within Document Discourse Patterns $\theta_{i,j}(x_i, x_j)$ and $\theta_i(x_i)$. We encode different types of knowledge through the pairwise and the singleton potentials. The pairwise potentials encode a number of discourse cues for the progress of information categories in the document: (1) Passive verbs tend to indicate category change; (2) Category change is highly likely in the opening part of a document; and (3) The closing part of a document is devoted mainly to reporting results and conclusions.

Therefore, if a sentence contains a passive verb or appears in the opening part[3] of a document, the pairwise potentials of that sentence and of its predecessor give a bonus to assignments in which a category change occur. Likewise, when moving from the opening part to the closing part of the document, the corresponding pairwise potentials encourage transition to a predefined set of clusters. [4]. The singleton potentials encode the tendency of scientific documents to start with a background knowledge related to the article, and to end with conclusions. They do so by encouraging the first sentence in each document to be in the first output category and, likewise, the last sentence in each document to be in the last output category.

Cross-Document Sentence Similarity. We build a feature representation for each vertex in X. We consider three of the feature sets described in (Guo and Korhonen, 2011): *POS* – the part-of-speech tags of the verbs in the sentence; *Location* – each document is divided into 10 parts, the location feature takes two values: the part where the sentence starts and the part where it ends; and *Object* – the words that appear as verb objects in the sentence .

We use this representation to encourage identical category assignment for similar sentences. We do this by two types of pairwise potentials: **(1) Pairwise potentials between sentence vertices ($\theta_{i,j}(x_i, x_j)$).** We define the similarity between the $i-th$ and the $j-th$ sentences, $SimScore_{i,j}$, as the number of features that have the same value in their representation. The similarity condition in the potential definition holds if the similarity score between the sentences is among the top M scores for the $i-th$ sentence; **(2) Pairwise potentials between sentence and agreement vertices ($\phi_{i,j}(x_i, a_j)$).** The similarity scores between sentences that belong to the same category tend to concentrate around the same value. Consequently a significant change in similarity between consecutive sentences is an indication of a category change. To encode this potential we scan the document from the beginning and compute the similarity between consecutive sentences. A similarity score that exceeds the maximum or deceeds the minimum of the previously observed similarity scores by a pre-defined threshold, is considered to be a class change indication [5]. For a sentence i that appears before the $j-th$ change we write $\phi(x_i, a_j) = \gamma$. The other values of this potential are set to zero. An example of two documents where a similarity pattern exists and of one document in which it does not exist (and therefore the ϕ values for its sentences with all agreement vertices are set to zero) is given in Figure 1.

[3] The opening part of a document is defined to be the first m^1 sentences, the rest of the sentences are considered to be its closing part. The average number of sentences in our abstracts is 10.3 and we set $m^1 = 4$.

[4] In our experiments we associated the last two clusters of the output scheme with the last part of the document as the AZ scheme we use for evaluation contains one cluster for results and one for conclusions.

[5] The maximum and minimum scores are computed over the sentence vertices from the previous change. For the first change the sentence vertices from the beginning of the document are considered.

Prior knowledge about Sentence Classification (Transduction) $\theta_i(x_i)$. We experiment with the conditions where we have oracle knowledge (i.e. knowledge that is taken from the gold standard) of the categories of some of the test set sentences and the model should predict the categories of the other sentences. The prior sentence-classification condition in the definition of $\theta_i(x_i)$ is simply that the category of the i-th sentence is known to be x_i.

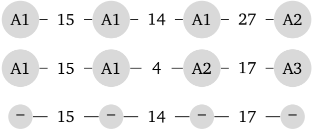
A1 – 15 – A1 – 14 – A1 – 27 – A2
A1 – 15 – A1 – 4 – A2 – 17 – A3
– — 15 — – — 14 — – — 17 — –

Figure 1: Three examples of similarity patterns in the beginning of documents. Lines represent documents, vertices represent sentences and the label inside a vertex corresponds to the agreement vertex to which this vertex is connected. Edges are labeled with the similarity score between the vertices they connect. The similarity difference threshold in this example is 10.

Inference Our model is a pairwise MRF. When cross-document sentence similarity knowledge is encoded, the model is very likely to have cycles which make exact inference NP-hard (see Section 3.1). When this knowledge is not encoded, the model becomes a simple linear chain model with edges between each pair of consecutive sentences in the same document. In such a model, exact inference can be done efficiently using dynamic programming. We addressed this problem by using the message passing algorithm for linear-programming (LP) relaxation of the MAP assignment (MPLP) described in (Sontag et al., 2008). LP relaxation algorithms for the MAP problem define an upper bound on the original objective which takes the form of a linear program. Consequently, a minimum of this upper bound can be found using standard LP solvers or, more efficiently, using specialised message passing algorithms (Yanover et al., 2006). The algorithm comes with an optimality guarantee: when the solution to the linear program is integral it is guaranteed to give the global optimum of the MAP problem. The MPLP algorithm described in (Sontag et al., 2008) is attractive in that it iteratively computes tighter upper bounds on the MAP problem.

4 Experiments

Data and Scenarios We experimented with the biomedical abstracts from the data set of (Guo et al., 2010) consisting of 1000 AZ-annotated abstracts (7985 sentences, 225785 words). We used the 792 abstract (7886 sentences) test set of (Guo and Korhonen, 2011). We consider two scenarios: a fully unsupervised scenario, and a transduction. For the latter we consider two conditions: (1) the identity of all the sentences that belong to one of the clusters is known; and (2) the oracle cluster assignment of randomly selected 5% or 10% of the sentences is known. In all cases our model as well as the baselines induce $K = 5$ categories.

Baselines Our first baseline is the K-means algorithm (Bishop, 2006) where sentences are represented by the same features that are used for constructing our similarity scores. In the fully unsupervised scenario we use the standard K-means. For transduction, in the condition where all the sentences of one of the classes are known, we run K-means only for the rest of the sentences and induce K-1 clusters; In the condition where the labels of a randomly selected sentence subset are known, we fix the classes of these sentences during the run of the algorithm (so that they affect their class centroid over the iterations) and use the mean of the vectors that are known to belong to each class as its initial centroid.

For the fully unsupervised scenario we also compare to the Hidden Topic Markov Model (HTMM) (Gruber et al., 2007) , a fully unsupervised, topic-model based algorithm. Like our within-document pairwise potentials, this algorithm models the sequential progress of topics in an abstract. However, in contrast to our model, it aims to maximize the lexical coherence of the induced categories. For the transduction scenario our second baseline is the transductive SVM algorithm (T-SVM from (Sinz, 2011)), where the feature-based representation is similar to the one we use in our model and in K-means. Being a classifier, this baseline is only useful in the second transduction condition where the categories of 5% or 10% of the sentences are given. T-SVM is a transductive classifier. To better understand the effect of each of these properties we also compare our model to the performance of a standard SVM . In addition to these baselines, we compare our full model to models created by omitting all the potentials related to a specific type of encoded knowledge: feature-based similarity or discourse patterns.

Parameter Tuning Our model governs the relative weight of its components with three potential parameters α, β and γ, and with M, the connectivity degree of the graph (Section 3). We manually set these parameters on 10 abstracts to the values that give the best performance in the unsupervised scenario and used them across all experiments. The potential parameters used are: $\alpha = 10^5$, $\beta = 10^2$, $\gamma = 10^7$ and M was set to 50.

We run K-means 100 times, randomly selecting the cluster centers from the set of clustered vectors, and selected the output clustering with the highest objective values. For HTMM, we assumed symmetric prior and ran the algorithm 10 times for each hyperparameter value in $\{0.01, 0.05, 0.1, 0.15 \ldots 1\}$. For each parameter assignment we selected the solution with the highest likelihood and the results we report are of oracle selection of the best of these solutions. The SVM algorithms were trained with the default setting of UniversSVM (Sinz, 2011).

Evaluation We uses greedy 1-1 mapping for evaluation. We mapped each induced category in the test set to one of the gold classes in a greedy 1-1 manner using the Kuhn-munkres algorithm for maximum matching in a bi-partite graph (Munkres, 1957). We then report the sentence level accuracy across the entire test set. In addition, we report per-class F-score (adjusted to a 0-100 scale) after the greedy 1-1 mapping is performed.

5 Results

The Fully Unsupervised Scenario Table 1 (left) presents results for the unsupervised setup. The top line shows results for the full test set: our model outperforms the K-means and HTMM baselines by 7% and 17.3%, respectively. The bottom lines present a similar pattern for the per-class F-score: our model is better for the BACKGROUND, RESULTS and CONCLUSIONS zones by up to 52%. This result demonstrates the importance of modelling linguistic similarity between sentences jointly with the sequential progression of the abstract discourse. While K-means clusters sentences together according to their linguistic similarity and HTMM models the progression of lexical topics in the abstract, our approach is to model linguistic similarity and sequential category progress jointly. Furthermore, unlike HTMM, we capitalize on discourse elements rather than on lexical cohesion when modelling the sequential progress.

The Transduction Scenario Results for this setup are in Table 2 (left). The first five rows correspond to the condition where all the sentences belonging to one of the gold classes are known. The SVM classifiers are not applicable to this condition as they cannot learn categories

	Results			Ablation Analysis		
Class	Full Model	K-means	HTMM	Full Model	Model - Similarity	Model - Disc.
All classes	**61.5**	54.5	44.2	**61.5**	58.5	53.0
Background (14.1%)	**58.4**	44.5	12.0	**58.4**	51.0	58.3
Objective (8.8%)	18.9	32.5	**33.0**	18.9	**22.7**	4.6
Methods (15.4%)	32.0	0	**46.0**	**32.0**	23.0	5.4
Results (45.3%)	**71.8**	66.6	57.6	**71.8**	70.2	58.3
Conclusions (15.4%)	**70.0**	50.8	18.0	**70.0**	64.3	57.5

Table 1: Results for the fully unsupervised scenario. Top line is for 1-1 accuracy for the full data set. Bottom lines are for per-class F-score.

	Results				Ablation Analysis		
Condition	Full Model	K-means	SVM	T-SVM	Full Model	Model - Similarity	Model - Disc.
Known background	**72.3**	65.6	—	—	**72.3**	70.7	58.5
known Obj.	**70.9**	63.3	—	—	**70.9**	66.5	59.4
Known Method	**77.6**	72.9	—	—	**77.6**	72.3	65.3
Known Results	**79.6**	74.5	—	—	79.6	72.7	**80.3**
Known Conclusion	**69.0**	63.2	—	—	**69.0**	64.5	67.9
Known 5% Sen.	**63.4**	59.0	59.9	54.2	**63.4**	61.2	52.5
Known 10% Sen.	**65.0**	60.5	61.9	55.8	**65.0**	63.2	56.5

Table 2: Results for the transduction scenario. In each of the first five lines the identity of all the sentences that belong to one of the classes is given to the models. In the last two lines the classes of a random sample of 5% or 10% of the sentences are known.

that do not appear in the training data. Our model is better than K-means in propagating this knowledge achieving 5.1% - 7.6% performance gain. The next two lines of the table compare the performance of the models when the categories of randomly selected 5% or 10% of the test-set sentences are known. Our model is superior again beating the baslines by 3.1% or more.

Model Components The right sections of the tables present an ablation analysis where we compare the performance of our full model to that of its components. When excluding the potentials that model between-document similarity from our model (Model - Similarity), the performance drops by 3% for the full test set (Table 1 top) and by up to 9% for four of the zone classes (Table 1 bottom) in the unsupervised scenario. Our full model further outperforms its discourse component in the seven transduction scenarios by up to 6.9%. When excluding these potentials from the model (Model - Discourse), the performance in the unsupervised scenario drops by 8.5% for the full test set and by up to 26.6% for the per-class F-score. Similarly, the performance drops in six of the seven transductive scenarios, by up to 13.8%.

Convergence The MPLP algorithm minimizes an upper bound on the MAP objective. Since this bound is convex, the MPLP algorithm is promised to converge to its global minimum, but the bound is promised to be tight only if the solution is integral – i.e. if every vertex is assigned to the same category by all the potentials that take it as an argument. In practice, in all the experimental conditions for all test subsets our model converges to an integral exact solution.

Conclusion and perspectives

We presented a novel unsupervised model for inferring information structure of scientific documents. The model integrates within-document discourse patterns and cross-document, feature-based linguistic information in a flexible way that enables to control the relative importance of different knowledge types by parameter setting. In the future we intend to extend our model to address more information sources and to use it for data-driven analysis of the various existing AZ schemes.

Acknowledgments

The work in this paper was funded by the Royal Society, (UK), EPSRC (UK) grant EP/G051070/1 and EU grant 7FP-ITC-248064.

References

Argamon, S., Dodick, J., and Chase, P. (2008). Language use reflects scientific methodology: A corpus-based study of peer reviewed journal articles. *Scientometrics*, 75(2):203–238.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, New-York.

Bunescu, R. and Mooney, R. (2004). Collective information extrction with relational markov networks. In *Proceedings of ACL*.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., and Bivens-Tatum, J. (2008). Cognitive models of writing: Writing proficiency as a complex integrated skill. *ETS RR-08-55*.

Finkel, J., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*.

Gammerman, A., Volk, V., and Vapnik, V. (1998). Learning by transduction. In *Proceedings of UAI*.

Ganchev, K., Graca, J., Gillenwater, J., and Taskar, B. (2010). Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Gillenwater, J., Ganchev, K., Graca, J., Pereira, F., and Taskar, B. (2010). Sparsity in dependency grammar induction. In *Proceedings of ACL Short Papers*.

Gruber, A., Rosen-Zvi, M., and Weiss, Y. (2007). The hidden topic markov model. In *Proceedings of AISTAT*.

Guo, Y. and Korhonen, A. (2011). A weakly supervised approach to argumantative zoning of scientific documents. In *Proceedings of EMNLP*.

Guo, Y., Korhonen, A., Liakata, M., Karolinska, I. S., Sun, L., and Stenius, U. (2010). Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*.

Gupta, R., Sarawagi, S., and Diwan, A. (2010). Collective inference for extraction mrfs coupled with symmetric clique potentials. *Journal of Machine Learning Research*.

Hachey, B. and Grover, C. (2006). Extractive summarisation of legal texts. *Artif. Intell. Law*, 14:305–345.

Hirohata, K., Okazaki, N., Anaiadou, S., and Ishikika, M. (2008). Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of IJCNLP*.

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Liang, P., Jordan, M., and Klein, D. (2009). Learning from measurements in exponential families. In *Proceedings ICML*.

Lin, J., Karakos, D., Demner-Fushman, D., and Khudanpur, S. (2006). Generative content models for structural analysis of medical abstracts. In *Proceedings of BioNLP-06*.

Lungen, H., Barenfanger, M., Hilbert, M., Lobin, H., and Puskas, C. (2010). Discourse relations and document structure. *Text, Speech and Language Technology*, 41:97–123.

Mann, G. and McCallum, A. (2010). Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11:955–984.

Mizuta, Y., Korhonen, A., Mullen, T., and Collier, N. (2006). Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics on Natural Language Processing in Biomedicine and Its Applications*, 75(6):468–487.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the SIAM*, 5(1):32–38.

Reichart, R. and Barzilay, R. (2012). Multi event extraction guided by global constraints. In *Proceedings of NAACL-HLT*.

Roth, D. and Yih, W. (2005). Integer linear programming inference for conditional random fields. In *Proceedings of ICML*.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbuhler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis, C., and Veuthey, A. L. (2007). Using argumentation to extract key sentences from biomedical abstracts. *Int J Med Inform*, 76(2-3):195–200.

Rush, A., Reichart, R., Collins, M., and Globerson, A. (2012). Improved parsing and pos tagging using inter-sentence consistency constraints. In *Proceedings of EMNLP*.

Shatkay, H., Pan, F., Rzhetsky, A., and Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

Sinz, F. (2011). *UniverSVM Support Vector Machine with Large Scale CCCP Functionality*. http://www.kyb.mpg.de/bs/people/fabee/universvm.html.

Sontag, D., Meltzer, T., Globerson, A., Jaakkola, T., and Weiss, Y. (2008). Tightening lp relxations for map using message passing. In *UAI*.

Sutton, C. and McCallum, A. (2004). Collective segmentation and labeling of distant entities in information extractions. In *ICML workshop on Statistical Relational Learning and its Applications*.

Taboada, M. and Mann, W. (2006). Applications of rhetorical structure theory. *Applications of Rhetorical Structure Theory*, 8(4):567–588.

Taskar, B., Abbeel, P., and Koller, D. (2002). Discriminative probabilistic models for relational data. In *Proceedings of UAI*.

Tbahriti, I., Chichester, C., Lisacek, F., and Ruch, P. (2006). Using argumentation to retrieve articles with similar citations. *Int J Med Inform*, 75(6):488–495.

Teufel, S. (2001). Task based evaluation of summary quality: Describing relationships between scientific papars. In *NAACL workshop on Automatic Text Summarization*.

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

Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP*.

Varge, A., Preotiuc-Pietro, D., and Ciravegna, F. (2012). Unsupervised document zone identification using probabilistic graphical models. In *Proceedings of LREC*.

Yanover, C., Meltzer, T., and Weiss, Y. (2006). Linear programming relazations and belief propogataion – and empitical study. *JMLR Special Issue on Machine Learning and Large Scale Optimization*.