Objective Function Learning to Match Human Judgements for Optimization-Based Summarization

Maxime Peyrard and **Iryna Gurevych**
Research Training Group AIPHES and UKP Lab
Computer Science Department, Technische Universität Darmstadt
www.aiphes.tu-darmstadt.de, www.ukp.tu-darmstadt.de

Abstract

Supervised summarization systems usually rely on supervision at the sentence or n-gram level provided by automatic metrics like ROUGE, which act as noisy proxies for human judgments. In this work, we learn a summary-level scoring function θ including human judgments as supervision and automatically generated data as regularization. We extract summaries with a genetic algorithm using θ as a fitness function. We observe strong and promising performances across datasets in both automatic and manual evaluation.

1 Introduction

The task of extractive summarization can naturally be cast as a discrete optimization problem where the text source is considered as a set of sentences and the summary is created by selecting an optimal subset of the sentences under a length constraint (McDonald, 2007). This view entails defining an objective function which is to be maximized by some optimization technique. In the ideal case, this objective function would encode all the relevant quality aspects of a summary, such that by maximizing all these quality aspects we would obtain the best possible summary.

However, we find several issues with the objective function in previous work on optimization-based summarization. First, the choice of the objective function is based on ad-hoc assumptions about which quality aspects of a summary are relevant (Kupiec et al., 1995). This bias can be mitigated via supervised techniques guided by data. In practice, these approaches use signals at the sentence (Conroy and O'leary, 2001; Cao et al., 2015) or n-gram (Hong and Nenkova, 2014; Li et al., 2013) level and then define a combination function to estimate the quality of the whole summary (Carbonell and Goldstein, 1998; Ren et al., 2016). This combination θ determines the trade-off between conflicting quality aspects (importance vs redundancy) encoded in the objective function by making simplistic assumptions to ensure convenient mathematical properties of θ like linearity or submodularity (Lin and Bilmes, 2011). This restriction comes from computational considerations without conceptual justifications. More importantly, the supervision signal comes from automatic metrics like ROUGE (Lin, 2004) which are convenient but noisy approximations for human judgment.

In this work, we propose to learn the objective function θ at the summary-level from a pool of manually annotated system summaries to ensure the extraction of summaries considered *good* by humans. This explicitly targets the extraction of high-quality summaries as measured by humans and limits undesired gaming of the target evaluation metric. However, the number of data points is relatively low and the learned θ might not be well-behaved (high θ scores for bad summaries) pushing the optimizer to explore regions of the feature space unseen during training where θ wrongly assumes high scores. To prevent this scenario, we rely on a large amount of noisy but automatic training data providing supervision on a larger span of the feature space. Intuitively, it can be viewed as a kind of regularization.

By defining θ directly at the summary-level, one has access to features like redundancy or global information content without the need to define a combination function from individual sentence scores. Any feature available at the sentence or n-gram level can be transferred to the summary-level (by summation), while the summary-level perspective provides access to new features capturing the interactions between sentences. Furthermore, recent works have demonstrated that global optimization using genetic algorithms without im-

Proceedings of NAACL-HLT 2018, pages 654–660
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

posing any mathematical restrictions on θ is feasible (Peyrard and Eckle-Kohler, 2016).

In summary, our contributions are: (1) We propose to learn a summary-level scoring function θ and use human judgments as supervision. (2) We demonstrate a simple regularization strategy based on automatic data generation to improve the behavior of θ under optimization. (3) We perform both automatic and manual evaluation of the extracted summaries, which indicate competitive performances.

2 Approach

2.1 Learning setup

Let θ^* be the observed human judgments. θ^* can be manual Pyramid (Nenkova et al., 2007) or overall responsiveness on a 0 to 5 LIKERT scale. We learn a function θ_w with parameters w approximating θ^* based on a feature set Φ. $\Phi(S) \in \mathbb{R}^d$ is the feature representation of a summary S.

Let $\mathcal{T}$ be the set of topics in the training set, and $\mathcal{S}_T$ the set of scored summaries for the topic T. The learning problem consists in minimizing the following loss function:

$$\mathcal{L}_\omega = \sum_{T \in \mathcal{T}} \sum_{s \in \mathcal{S}_T} ||\theta_\omega(\Phi(S)) - \theta^*(S)||^2 \qquad (1)$$

While any regression algorithm could be applied, we observed strong performances for the simple linear regression. It is particularly simple and not prone to overfitting.

2.2 Automatic data generation

Few annotated summaries are available (50 per topic) and they cover a small region of the feature space (low variability). θ may wrongly assume high scores in some parts of the feature space despite lack of evidence. The optimizer will explore these regions and output low-quality summaries.

To address this issue, we generate summaries distributed across the feature space. For each feature x, we sample a set of $k = 100$ summaries covering the range of possible values of x. For sampling, we use the genetic algorithm recently introduced by Peyrard and Eckle-Kohler (2016).[1] Their solver implements a Genetic Algorithm (**GA**) to create and iteratively optimize summaries over time. We use default values for the reproduction and mutation rate and set the population size to 50. With x as fitness function, the resulting population is a set of summaries ranging from random to (close to) maximal value. After both maximization and minimization, we obtain 100 summaries covering the full range of x.

In total, we sample $m \cdot k$ summaries per topic, where m is the number of features. We score these summaries with ROUGE-2 recall (R2), which is a noisy approximation of human judgments but provides indications preventing bad regions from getting high scores.

2.3 Summary Extraction

We trained 3 different scoring functions: θ_{pyr} with manual pyramid annotations; θ_{resp} with responsiveness annotations; and θ_{R2} with our automatically generated data. [2] The final scoring function is a linear combination:

$$\theta(S) = \alpha_1 \cdot \theta_{pyr}(S) + \alpha_2 \cdot \theta_{resp}(S) + \alpha_3 \cdot \theta_{R2}(S)$$

Therefore θ_{R2} acts as a regularizer for the θ's learned with human judgments. [3] It is a simple form of model averaging which combine the different information of the 3 different models.

We didn't constrain θ to have specific properties like linearity with respect to sentence scores, thus extracting high scoring summaries cannot be done with Integer Linear Programming. Instead, we search an approximate solution by employing the same meta-heuristic solver we used for sampling with θ as the fitness function.

2.4 Features

Learning a scoring function at the summary-level gives us access to both n-gram/sentence-level features and summary-level features. Sentence-level features can be transferred to the summary-level, while new features capturing the interactions between sentences in the summary become available.

As sentence-level features, we used the standard: TF*IDF, n-gram frequency and overlap with the title. As new summary-level features, we used: number of sentences, summary-level redundancy and summary-level n-gram distributions: Jensen-Shannon (JS) divergence with n-gram distribution in the source (Louis and Nenkova, 2013).

[1] https://github.com/UKPLab/coling2016-genetic-swarm-MDS

[2] We train these models separately because the different annotations do not lie on the same scale

[3] We didn't automatically tune the different values of α but observed that $[1, 0.5, 0.5]$ works well in practice.

N-gram Coverage. Each n-gram g_i in the documents has a frequency $tf(g_i)$, the summary S is scored by:

$$Cov_n(S) = \sum_{g \in S_n} tf(g_i)$$

Here S_n is the multiset of n-grams (with repetitions) composing S. Also, the frequency can be computed either by counting the number of occurrence of the n-gram or by counting the number of documents in which the n-gram appears. For both frequency computations, we extract features for unigrams, bigrams and trigrams.

TF*IDF. Each n-gram g_i is also associated its Inverse Document Frequence: $idf(g_i)$. The summary S is scored by:

$$TF * IDF_n(S) = \sum_{g \in S_n} tf(g_i) * idf(g_i)$$

Here S_n is the multiset of n-grams (with repetitions) composing the summary S. We also extract features for both frequency computations for unigrams, bigrams and trigrams.

Overlap with title. We measure the proportion of n-grams from the title that appear in the summary:

$$Overlap_n(S) = \frac{|T_n \cap S_n|}{T_n}$$

Where T_n is the multiset of n-grams in the title, and S_n is the multiset of n-grams in the summary. We compute it for unigrams, bigrams and trigrams.

Number of sentences. We also use the number of sentences in S as a feature because summaries with a lot of sentences tend to have very short and meaningless sentences.

Redundancy. Previous features were at the sentence-level, we obtained features for the whole summary by summation over sentences. However, the redundancy of S cannot be computed at the sentence-level. This is an example of features available at the summary-level but not available at the sentence-level. We define it as the number of unique n-gram types ($|U_n|$) in the summary divided by the total number of n-gram tokens (the length of S)

$$Red_n(S) = \frac{|U_n|}{|S_n|}$$

Where U_n is the set of n-grams (without repetitions) composing S and S_n is the multiset of n-grams (with repetitions).

Divergences. This is another feature that can only be computed at the summary-level inspired by Haghighi and Vanderwende (2009) and Peyrard and Eckle-Kohler (2016). We compute the KL divergence and JS divergence between n-gram probability distributions of the summaries and of the documents. The probability distributions are built from the two kinds of frequency distributions and for unigrams, bigrams and trigrams.

3 Experiments

Dataset We use two multi-document summarization datasets from the Text Analysis Conference (TAC) shared tasks: TAC-2008 and TAC-2009.[4] TAC-2008 and TAC-2009 contain 48 and 44 topics, respectively. Each topic consists of 10 news articles to be summarized in a maximum of 100 words. We use only the so-called initial summaries (A summaries), but not the update part.

We used these datasets because all system summaries and the 4 reference summaries were manually evaluated by NIST assessors for content selection (with Pyramid) and overall responsiveness. At the time of the shared tasks, 57 systems were submitted to TAC-2008 and 55 to TAC-2009. For our experiments, we use the Pyramid and the responsiveness annotations.

With our notations, for example with TAC-2009, we have $n = 55$ scored system summaries, $m = 44$ topics, $\mathcal{D}_i$ contains 10 documents and θ_i contains 4 reference summaries.

We also use the recently created German dataset DBS (Benikova et al., 2016) which contains 10 heterogeneous topics. For each topic, 5 summaries were evaluated by trained human annotators but only for content selection with Pyramid. The summaries have variable sizes and are about 500 words long.

Baselines (1) ICSI (Gillick and Favre, 2009) is a global linear optimization approach that extracts a summary by solving a maximum coverage problem considering the most frequent bigrams in the source documents. ICSI has been among the best systems in a standard ROUGE evaluation (Hong et al., 2014). (2) LexRank (Erkan

[4] http://tac.nist.gov/2009/Summarization/, http://tac.nist.gov/2008/

	ρ	NDCG
Best-Baseline-R	.594	.505
θ_{R2}	**.663**	**.536**
Best-Baseline-Pyr	.492	.715
θ_{pyr}	**.554**	**.780**
Best-Baseline-Resp	.367	.710
θ_{resp}	**.391**	**.741**

Table 1: Performance of learned θ's compared to the best baselines for each type annotation types.

and Radev, 2004) is a graph-based approach computing sentence centrality based on the PageRank algorithm. (3) KL-Greedy (Haghighi and Vanderwende, 2009) minimizes the Kullback-Leibler (KL) divergence between the word distributions in the summary and the documents. (3) Peyrard and Eckle-Kohler (2016) optimize JS divergence with a genetic algorithm. (4) Finally, SFOUR is a supervised structured prediction approach that trains an end-to-end on a convex relaxation of ROUGE (Sipos et al., 2012).

Objective function learning In this section, we measure how well our models can predict human judgments. We train each θ in a leave-one-out cross-validation setup for each dataset and compare their performance to the summary scoring function of baselines like it was done previously (Peyrard and Eckle-Kohler, 2017). Each individual feature is also included in the baselines.

Correlations are measured with two complementary metrics: Spearman's ρ and Normalized Discounted Cumulative Gain (NDCG). Spearman's ρ is a rank correlation metric, which compares the ordering of systems induced by θ and the ordering of systems induced by human judgments. NDCG is a metric that compares ranked lists and puts more emphasis on the top elements with logarithmic decay weighting. Intuitively, it captures how well θ can recognize the best summaries. The optimization scenario benefits from high NDCG scores because only summaries with high θ scores are extracted.

The results are presented in Table 1. For simplicity, we report the average over the 3 datasets. Each θ is compared against the best performing baseline for the data annotation type it was trained on (R2, responsiveness or pyramid).[5] The trained models perform substantially and consistently better than the best baselines. They have a high correlation with human judgments and are capable of identifying *good* summaries.

However, we need to test whether the combination of the three θ's is well behaved under optimization. For this, we perform an evaluation of the summaries extracted by the genetic optimizer.

Summaries Evaluation Now, we evaluate the summaries extracted by the genetic optimizer with θ as fitness function (noted (θ, Gen)). We still train θ with leave-one-out cross-validation.

To evaluate summaries, we report the ROUGE variant identified by Owczarzak et al. (2012) as strongly correlating with human evaluation methods: ROUGE-2 (R2) recall with stemming and stopwords not removed. We also report JS2, the Jensen-Shannon divergence between bigrams in the reference summaries and the candidate system summary (Lin et al., 2006). The last metric is S3 (Peyrard et al., 2017), a combination of several existing metrics trained explicitly to maximize its correlation with human judgments.

Finally, our approach aims at improving summarization systems based on human judgments, therefore we also set up a manual evaluation for the two English datasets. Two annotators were given the summaries of every system for 10 randomly selected topic of both TAC-2008 and TAC-2009. They annotated (with a Cohen's kappa of 0.73) summaries on a LIKERT scale following the responsiveness guidelines.

The results are reported in Table 2. We perform significance testing with *Approximate Random Testing* to compare differences between two means in cross-validation [6].

While θ's trained on human judgments have a high correlation with human judgments, they behave badly under optimization. This effect is much less visible for θ_{R2} because the data points have been sampled to cover the feature space. We observe the effectiveness of the regularization because each $\theta_{R2/pyr/resp}$ performs much worse individually than the combined θ. We also note that (θ_{R2}, Gen) performs on par with the other supervised baseline SFOUR but both are outperformed by exploiting human judgments. (θ, Gen) is consistently and often significantly better than baselines across datasets and metrics. In particular, humans tend to prefer the summaries extracted by

[5] Best baseline for R2 and Responsiveness is: KL divergence on bigrams; for Pyramid: KL divergence on unigrams

[6] The symbol * indicates that the difference compared to the previous best baseline is significant with $p \leq 0.05$

	TAC-2008				TAC-2009				DBS		
	R2↑	JS2↓	S3↑	H↑	R2↑	JS2↓	S3↑	H↑	R2↑	JS2↓	S3↑
LexRank	.078	.635	.336	3.74	.090	.625	.360	3.75	.105	.594	.354
(KL, Greedy)	.068	.644	.294	3.42	.061	.648	.288	3.21	.078	.620	.293
(JS, Gen)	.098	.618	.376	3.99	.101	.618	.370	3.89	.112	**.584**	.362
SFOUR	.101	.623	.372	3.88	.101	.622	.367	3.85	.114	.591	.357
ICSI	.101	.620	.377	4.03	.103	.619	.369	3.91	.115	.586	.361
(θ_{R2}, Gen)	.100	.620	.375	3.89	**.104**	.618	.373	3.82	.116	.585	.363
(θ_{pyr}, Gen)	.096	.623	.369	3.65	.085	.631	.339	3.77	.078	.615	.312
(θ_{resp}, Gen)	.096	.622	.364	3.78	.085	.635	.342	3.88	-	-	-
(θ, Gen)	**.105**	**.615***	**.382**	**4.09***	**.104**	**.617**	**.376**	**4.03***	**.117**	**.584**	**.367***

Table 2: Comparison of systems across 3 datasets evaluated with ROUGE-2 recall; JS divergence on bigrams; S3 and Human annotations.

(θ, Gen). Manual inspection of summaries reveals that (θ, Gen) has lower redundancy than previous baselines thanks to summary-level features.

Important Features Since we used a linear regression, we can estimate the contribution of a feature by the amplitude of its associated weight. The two best features (n-gram distributions and redundancy) are summary-level features, which confirms the advantage of using a summary-level scoring function.

4 Related Work and Discussion

Supervised summarization started with Kupiec et al. (1995) who observed that there is no principled method to select and weight relevant features. Previous work focused on predicting sentence (Conroy and O'leary, 2001; Cao et al., 2015) or n-gram (Hong and Nenkova, 2014; Li et al., 2013) scores and then defining a composition function to get a score for the summary. This combination usually accounts for redundancy or coherence (Nishikawa et al., 2014) in an ad-hoc fashion (Carbonell and Goldstein, 1998; Ren et al., 2016). Structure prediction has been investigated to learn the composition function as well (Sipos et al., 2012; Takamura and Okumura, 2010). The supervision is always provided by automatic metrics, whereas we incorporate human judgments as supervision and learn from it directly at the summary-level. We note that He et al. (2006) and Peyrard and Eckle-Kohler (2016) have used a scoring function at the summary-level but these approaches are unsupervised.

One of the challenges we face is the lack of data with human judgments. We hope that this work will encourage efforts to create new and large datasets as they will be decisive for the progress of summarization. Indeed, systems trained only with automatic metrics can only be as good as the metrics are as a proxy for humans.

We used simple features but using more complex and semantic features is promising. Indeed, two syntactically similar but semantically different summaries cannot be distinguished by ROUGE, which diminishes the usefulness of semantic features. However, humans can distinguish them, thus inducing better usage of such features.

Another promising direction is to investigate more sophisticated ways of combining the human judgments with the automatically generated data. For example, by exploiting techniques from semi-supervised learning (Zhu et al., 2009) or by dynamically sampling unseen regions of the feature space with active learning (Settles, 2009).

5 Conclusion

We proposed an approach to learn a summary-level scoring function θ with human judgments as supervision and automatically generated data as regularization. The summaries subsequently extracted with a genetic algorithm are of high quality according to both automatic and manual evaluation. We hope this work will encourage more research directed towards the generation and usage of human judgment datasets.

Acknowledgements

This work has been supported by the German Research Foundation (DFG) as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1, and via the German-Israeli Project Cooperation (DIP, grant No. GU 798/17-1).

References

Darina Benikova, Margot Mieskes, Christian M. Meyer, and Iryna Gurevych. 2016. Bridging the Gap Between Extractive and Abstractive Summaries: Creation and Evaluation of Coherent Extracts from Heterogeneous Sources. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1039 – 1050.

Ziqiang Cao, Furu Wei, Sujian Li, Wenjie Li, Ming Zhou, and Houfeng Wang. 2015. Learning summary prior representation for extractive summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 829–833.

Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

John M. Conroy and Dianne P. O'leary. 2001. Text summarization via hidden markov models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 406–407.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *Journal of Artificial Intelligence Research*, pages 457–479.

Dan Gillick and Benoit Favre. 2009. A Scalable Global Model for Summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing (ILP'09)*, pages 10–18.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-document Summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Yan-xiang He, De-xi Liu, Dong-hong Ji, Hua Yang, and Chong Teng. 2006. MSBGA: A Multi-Document Summarization System Based on Genetic Algorithm. In *2006 International Conference on Machine Learning and Cybernetics*, pages 2659–2664.

Kai Hong, John Conroy, benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616.

Kai Hong and Ani Nenkova. 2014. Improving the Estimation of Word Importance for News Multi-Document Summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 712–721.

Julian Kupiec, Jan Pedersen, and Francine 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.

Chen Li, Xian Qian, and Yang Liu. 2013. Using Supervised Bigram-based ILP for Extractive Summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1004–1013.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of ACL workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An Information-Theoretic Approach to Automatic Evaluation of Summaries. In *Proceedings of the Human Language Technology Conference at NAACL*, pages 463–470, New York City, USA.

Hui Lin and Jeff A. Bilmes. 2011. A Class of Submodular Functions for Document Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 510–520, Portland, Oregon.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):267–300.

Ryan McDonald. 2007. A Study of Global Inference Algorithms in Multi-document Summarization. In *Proceedings of the 29th European Conference on Information Retrieval Research*, pages 557–564.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transaction on Speech and Language Processing*, 4.

Hitoshi Nishikawa, Kazuho Arita, Katsumi Tanaka, Tsutomu Hirao, Toshiro Makino, and Yoshihiro Matsuo. 2014. Learning to generate coherent summary with discriminative hidden semi-markov model. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1648–1659.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to Score System Summaries for Better Content Selection Evaluation. In *Proceedings of the EMNLP workshop New Frontiers in Summarization*, pages 74–84.

Maxime Peyrard and Judith Eckle-Kohler. 2016. A General Optimization Framework for Multi-Document Summarization Using Genetic Algorithms and Swarm Intelligence. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 247 – 257.

Maxime Peyrard and Judith Eckle-Kohler. 2017. A Principled Framework for Evaluating Summarizers: Comparing Models of Summary Quality against Human Judgments. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, volume Volume 2: Short Papers. Association for Computational Linguistics.

Pengjie Ren, Furu Wei, Zhumin Chen, Jun Ma, and Ming Zhou. 2016. A redundancy-aware sentence regression framework for extractive summarization. In *26th International Conference on Computational Linguistics, Proceedings of the Conference*, pages 33–43.

Burr Settles. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin Learning of Submodular Summarization Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 224–233.

Hiroya Takamura and Manabu Okumura. 2010. Learning to Generate Summary as Structured Output. In *Proceedings of the 19th ACM international Conference on Information and Knowledge Management*, pages 1437–1440.

Xiaojin Zhu, Andrew B. Goldberg, Ronald Brachman, and Thomas Dietterich. 2009. *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers.