Narrowing the Loop: Integration of Resources and Linguistic Dataset Development with Interactive Machine Learning

Seid Muhie Yimam

FG Language Technology
Department of Computer Science
Technische Universität Darmstadt
http://www.lt.tu-darmstadt.de
yimam@lt.informatik.tu-darmstadt.de

Abstract

This thesis proposal sheds light on the role of interactive machine learning and implicit user feedback for manual annotation tasks and semantic writing aid applications. First we focus on the cost-effective annotation of training data using an interactive machine learning approach by conducting an experiment for sequence tagging of German named entity recognition. To show the effectiveness of the approach, we further carry out a sequence tagging task on Amharic part-of-speech and are able to significantly reduce time used for annotation. The second research direction is to systematically integrate different NLP resources for our new semantic writing aid tool using again an interactive machine learning approach to provide contextual paraphrase suggestions. We develop a baseline system where three lexical resources are combined to provide paraphrasing in context and show that combining resources is a promising direction.

1 Introduction

Machine learning applications require considerable amounts of annotated data in order to achieve a good prediction performance (Pustejovsky and Stubbs, 2012). Nevertheless, the development of such annotated data is labor-intensive and requires a certain degree of human expertise. Also, such annotated data produced by expert annotators has limitations, such as 1) it usually does not scale very well since annotation of a very large data set is prohibitively expensive, and 2) for applications which should reflect dynamic changes of data over time, static training data will not serve its purpose. This issue is commonly known as *concept drift* (Kulesza et al., 2014).

There has been a lot of effort in automatically expanding training data and lexical resources using different techniques. One approach is the use of active learning (Settles et al., 2008) which aims at reducing the amount of labeled training data required by selecting most informative data to be annotated. For example it selects the instances from the training dataset about which the machine learning model is least certain how to label (Krithara et al., 2006; Settles, 2010; Raghavan et al., 2006; Mozafariy et al., 2012). Another recent approach to alleviate bottleneck in collecting training data is the usage of crowdsourcing services (Snow et al., 2008; Costa et al., 2011) to collect large amount of annotations from non-expert crowds at comparably low cost.

In an interactive machine learning approach, the application might start with minimal or no training data. During runtime, the user provides simple feedback to the machine learning process interactively by correcting suggestions or adding new annotations and integrating background knowledge into the modeling stage (Ware et al., 2002).

Similarly, natural language processing (NLP) tasks, such as information retrieval, word sense disambiguation, sentiment analysis and question answering require comprehensive external knowledge sources (electronic dictionaries, ontologies, or thesauri) in order to attain a satisfactory performance (Navigli, 2009). Lexical resources such as WordNet, Wordnik, and SUMO (Niles and Pease, 2001) also suffer from the same limitations that the machine learning training data faces.
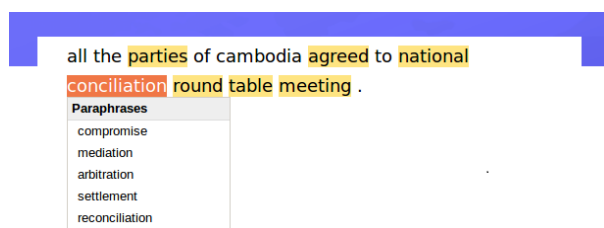
Figure 1: An online interface for the semantic writing aid application. Paraphrase suggestions are presented from a systematic combination of different NLP resources.

This proposal focuses on the development and enhancement of training data as well as on systematic combinations of different NLP resources for a semantic writing aid application. More specifically we address the following issues: 1) How can we produce annotated data of high quality using an interactive machine learning approach? 2) How can we systematically integrate different NLP resources? 3) How can we integrate user interaction and feedback into the interactive machine learning system? Moreover, we will explore the different paradigms of interactions (when should the machine learning produce a new model, how to provide useful suggestions to users, and how to control annotators behavior in the automation process). To tackle these problems, we will look at two applications, 1) an annotation task using a web-based annotation tool and 2) a semantic writing aid application, a tool with an online interface that provides users with paraphrase detection and prediction capability for a varying writing style. In principle, the two applications have similar nature except that the ultimate goal of the annotation task is to produce a fully annotated data whereas the semantic writing aid will use the improved classifier model instantly. We have identified a sequence tagging and a paraphrasing setup to explore the aforementioned applications.

Sequence tagging setup: We will employ an annotation tool similar to WebAnno (Yimam et al., 2014) in order to facilitate the automatic acquisition of training data for machine learning applications. Our goal is to fully annotate documents sequentially but interactively using the machine learning support in contrast to an active learning setup where the system presents portions of the document at a time.

Paraphrasing setup: The semantic writing aid

tool is envisioned to improve readability of documents and provide varied writing styles by suggesting semantically equivalent paraphrases and remove redundant or overused words or phrases. Using several lexical resources, the system will detect and provide alternative contextual paraphrases as shown in Figure 1. Such paraphrasing will substitute words or phrases in context with appropriate synonyms when they form valid collocations with the surrounding words (Bolshakov and Gelbukh, 2004) based on the lexical resource suggestion or using statistics gathered from large corpora. While the work of Bhagat and Hovy (2013) shows that there are different approaches of paraphrasing or quasi-paraphrasing based on syntactical analysis, we will also further explore context-aware paraphrasing using distributional semantics (Biemann and Riedl, 2013) and machine learning classifiers for contextual similarity.

2 Related Work

There have been many efforts in the development of systems using an adaptive machine learning process. Judah et al. (2009) developed a system where the machine learning and prediction process incorporates user interaction. For example, for sensitive email detection system, the user is given the opportunity to indicate which features, such as body or title of the message, or list of participants, are important for prediction so that the system will accordingly learn the classification model based on the user preference. Similarly, recommender systems usually provide personalized suggestions of products to consumers (Desrosiers and Karypis, 2011). The recommendation problem is similar to an annotation task as both of them try to predict the correct suggestions based on the existing user preference.

CueFlik, a system developed to support Web image search (Amershi et al., 2011), demonstrates that active user interactions can significantly impact the effectiveness of the interactive machine learning process. In this system, users interactively define visual concepts of pictures such as product photos or pictures with quiet scenery, and they train the system so as to learn and re-rank web image search results.

JAAB (Kabra et al., 2013) is an interactive machine learning system that allows biologists to use machine learning in closed loop without assistance

from machine learning experts to quickly train classifiers for animal behavior. The system allows users to start the annotation process with trustworthy examples and train an initial classifier model. Furthermore, the system enables users to correct suggestions and annotate unlabeled data that is leveraged in subsequent iteration.

Stumpf et al. (2007) investigate the impact of user feedback on a machine learning system. In addition to simple user feedback such as accepting and rejecting predictions, complex feedback like selecting the best features, suggestions for the reweighting of features, proposing new features and combining features significantly improve the system.

2.1 Combination and Generation of Resources

There are different approaches of using existing NLP resources for an application. Our approach mainly focuses on a systematic combination of NLP resources for a specific application with the help of interactive machine learning. As a side product, we plan to generate an application-specific NLP resource that can be iteratively enhanced.

The work by Lavelli et al. (2002) explores how thematic lexical resources can be built using an iterative process of learning previously unknown associations between terms and themes. The research is inspired by text categorization. The process starts with minimal manually developed lexicons and learns new thematic lexicons from the user interaction.

Jonnalagadda et al. (2012) demonstrate the use of semi-supervised machine learning to build medical semantic lexicons. They demonstrated that a distributional semantic method can be used to increase the lexicon size using a large set of unannotated texts.

The research conducted by Sinha and Mihalcea (2009) concludes that a combination of several lexical resources generates better sets of candidate synonyms where results significantly exceed the performance obtained with one lexical resource.

While most of the existing approaches such as UBY (Gurevych et al., 2012) strive at the construction of a unified resource from several lexical resources, our approach focuses on a dynamic and interactive approach of resource integration. Our approach is adaptive in such a way that the resource integration depends on the nature of the application.

3 Overview of the Problem

3.1 Interactive Machine Learning Approach

The generation of large amounts of high quality training data to train or validate a machine learning system at one pass is very difficult and even undesirable (Vidulin et al., 2014). Instead, an interactive machine learning approach is more appropriate in order to adapt the machine learning model iteratively using the train, learn, and evaluate technique.

Acquiring new knowledge from newly added training data on top of an existing trained machine learning model is important for incremental learning (Wen and Lu, 2007). An important aspect of such incremental and interactive machine learning approach is, that the system can start with minimal or no annotated training data and continuously presents documents to a user for annotation. On the way, the system can learn important features from the annotated instances and improve the machine learning model continuously. When a project requires to annotate the whole dataset, an interactive machine learning approach can be employed to incrementally improve the machine learning model.

3.2 Paraphrasing and Semantic Writing Aid

Acquisition and utilization of contextual paraphrases in a semantic writing aid ranges from integration of structured data sources such as ontologies, thesauri, dictionaries, and wordnets over semi-structured data sources such as Wikipedia and encyclopedia entries to resources based on unstructured data such as distributional thesauri. Paraphrases using ontologies such as YAGO (Suchanek et al., 2007) and SUMO provide particular semantic relations between lexical units. This approach is domain specific and limited to some predefined form of semantic relations. Structured data sources such as WordNet support paraphrase suggestions in the form of synonyms. Structured data sources have limited coverage and they usually do not capture contextual paraphrases. Paraphrases from unstructured sources can be collected using distributional similarity techniques from large corpora. We can also obtain paraphrase suggestions from monolingual comparable corpora, for example, using multiple translations of foreign novels (Ibrahim et al., 2003) or different news articles about the same topics (Wang

and Callison-Burch, 2011). Moreover, paraphrases can also be extracted from bilingual parallel corpora by "pivoting" a shared translation and ranking paraphrases using the translation probabilities from the parallel text (Ganitkevitch and Callison-Burch, 2014).

The research problem on the one hand is the adaptation of such diverse resources on the target semantic writing aid application and on the other hand the combination of several such resources using interactive machine learning to suit the application.

4   Methodology: Paraphrasing Component

The combinations of lexical resources will be based on the approach of Sinha and Mihalcea (2009), where candidate synonymous from different resources are systematically combined in a machine learning framework. Furthermore, lexical resources induced in a data driven way such as distributional thesauri (DT) (Weeds and Weir, 2005), will be combined with the structured lexical resources in an interactive machine learning approach, which incrementally learns weights through a classifier. We will train a classifier model using features from resources, such as n-gram frequencies, co-occurrence statistics, number of senses from WordNet, different feature values from the paraphrase database (PPDB)[1] (Ganitkevitch and Callison-Burch, 2014), and syntactic features such as part of speech and dependency patterns. Training data will be acquired with crowdsourcing by 1) using existing crowdsourcing frameworks and 2) using an online interface specifically developed as a semantic writing aid tool (ref Figure 1).

While the way the system provides suggestions might be based on many possible conditions, we will particularly address at least the following ones: 1) non-fitting word detection, 2) detection of too many repetitions, and 3) detection of stylistic deviations.

Once we have the resource combining component in place, we employ an interactive machine learning to train a classifier based on implicit user feedback obtained as 1) users intentionally request paraphrasing and observe their actions (such as which of the suggestion they accept, if they ignore all suggestions, if the users provide new paraphrase by themselves, and so on), and 2) the system automatically suggests candidate paraphrases (as shown in Figure 1) and observe how the user interacts.

5   Experiments and Evaluation

We now describe several experimental setups that evaluate the effectiveness of our current system, the quality of training data obtained, and user satisfaction in using the system. We have already conducted some preliminary experiments and simulated evaluations towards some of the tasks.

5.1   Annotation Task

As a preliminary experiment, we have conducted an interactive machine learning simulation to investigate the effectiveness of this approach for named entity annotation and POS tagging tasks. For the named entity annotation task, we have used the training and development dataset from the GermEval 2014 Named Entity Recognition Shared Task (Benikova et al., 2014) and the online machine learning tool MIRA[2] (Crammer and Singer, 2003). The training dataset is divided by an increasing size, as shown in Table 1 to train the system where every larger partition contains sentences from earlier parts. From Figure 2 it is evident that the interactive machine learning approach improves the performance of the system (increase in recall) as users continue correcting the suggestions provided.

Sentences	precision	recall	F-score
24	80.65	1.12	2.21
60	62.08	6.68	12.07
425	71.57	35.13	47.13
696	70.36	43.02	53.40
1264	71.35	47.15	56.78
5685	77.22	56.57	65.30
8770	77.83	60.16	67.86
10 812	78.06	62.72	69.55
15 460	78.14	64.96	70.95
24 000	80.15	68.82	74.05

Table 1: Evaluation result for the German named entity recognition task using an interactive online learning approach with different sizes of training dataset tested on the fixed development dataset.

---

[1]http://paraphrase.org

[2]https://code.google.com/p/miralium/

Furthermore, an automation experiment is carried out for Amharic POS tagging to explore if interactive machine learning reduces annotation time. In this experiment, a total of 34 sentences are manually annotated, simulating different levels of precision and recall (ref Table 2) for automatic suggestions as shown in Figure 3. We have conducted this annotation task several times to measure the savings in time when using automatic annotation. When no suggestion is provided, it took about 67 minutes for an expert annotator to completely annotate the document. In contrast to this, the same annotation task with suggestions (e.g with recall of 70% and precision of 60%) took only 21 minutes, demonstrating a significant reduction in annotation cost.

		recall (%)			
		no Auto.	30	50	70
prec (%)	no Auto.	67	-	-	-
	60	-	53	33	21
	70	-	45	29	20
	80	-	42	28	18

Table 2: Experimentation of interactive machine learning for different precision and recall levels for Amharic POS tagging task. The cell with the precision/recall intersection records the total time (in minutes) required to fully annotate the dataset with the help of interactive automation. Without automation (no Auto.), annotation of all sentences took 67 minutes.
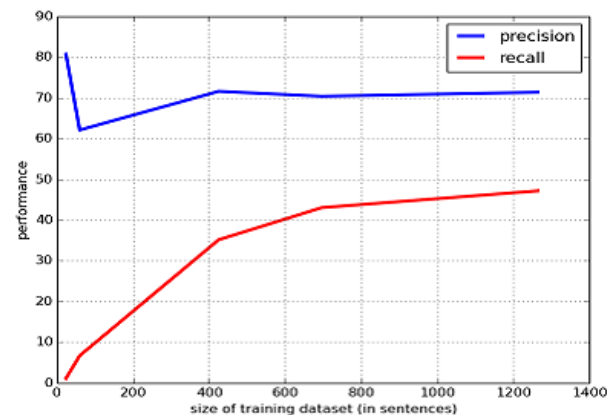

Figure 2: Learning curve showing the performance of interactive automation using different sizes of training data

5.2 Evaluation of Paraphrasing

For the semantic writing aid tool, we need to create a paraphrasing component (see Sec. 3.2). We conduct an evaluation by comparing automatic paraphrases against existing paraphrase corpora (Callison-Burch et al., 2008). The Microsoft Research Paraphrase Corpus (MSRPC) (Dolan et al., 2004) dataset, PPDB, and the DIRT paraphrase collections (Lin and Pantel, 2001) will be used for phrase-level evaluations. The TWSI dataset (Biemann, 2012) will be used for the word level paraphrase evaluation. We will use precision, recall, and machine translation metrics BLEU for evaluation.

Once the basic paraphrasing system is in place and evaluated, the next step will be the improvement of the paraphrasing system using syntagmatic and paradigmatic structures of language as features. The process will incorporate the implementation of distributional similarity based on syntactic structures such as POS tagging, dependency parsing, token n-grams, and patterns, resulting in a context-aware paraphrasing system, which offers paraphrases in context. Furthermore, interactive machine learning can be employed to train a model that can be used to provide context-dependent paraphrasing.

5.2.1 Preliminary Experiments

We have conducted preliminary experiments for a semantic writing aid system, employing the LanguageTools (Naber, 2004) user interface to display paraphrase suggestions. We have used WordNet, PPDB, and JobimText DT[3] to provide paraphrase
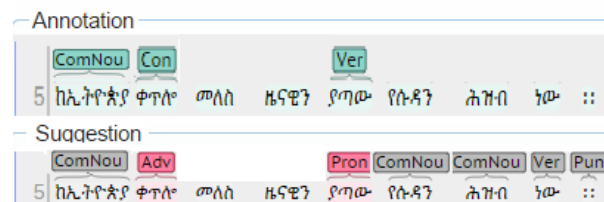
---

[3]http://goo.gl/0Z2Rcs

Figure 3: Amharic POS tagging. lower pane: suggestion provided to the user by the interactive classifier, upper pane: annotations by the user. When (grey) the suggestion in the lower pane is correct, the user will click the annotation and copy it to the upper pane. Otherwise (shown in red or no suggestion), the user should provide a new annotation in the upper pane.

suggestions. Paraphrases are first obtained from each individual resources and irrelevant or out-of-context paraphrases are discarded by ranking alternatives using an n-gram language model. Paraphrases suggested by most of the underlining resources (at least 2 out of 3) are provided as suggestions. Figure 1 shows an online interface displaying paraphrase suggestions based on our approach[4].

We have conducted experimental evaluation to assess the performance of the system using recall as a metric ($recall = \frac{s}{r}$ where s is the number of tokens in the source (paraphrased) sentence and r is the number of tokens in the reference sentence). We have used 100 sentences of paraphrase pairs (source and reference sentences) from the MSRPC dataset. The baseline result is computed using the original paraphrase pairs of sentences which gives us a recall of 59%. We took the source sentence and applied our paraphrasing technique for words that are not in the reference sentence and computed recall. Table 3 shows results for different settings, such as taking the first, top 5, and top 10 suggestions from the candidate paraphrases which outperforms the baseline result. The combination of different resources improves the performance of the paraphrasing system.

setups	Baseline	top 1	top 5	top 10
WordNet	59.0	60.3	61.4	61.9
ppdb	59.0	60.2	62.2	64.6
JoBimText	59.0	59.9	60.3	60.4
2in3	59.0	**60.7**	**65.3**	**66.2**

Table 3: Recall values for paraphrasing using different NLP resources and techniques. Top 1 is where we consider only the best suggestion and compute the score. top 5 and 10 considers the Top 5 and 10 suggestions provided by the system respectively. The row 2in3 shows the result where we consider a paraphrase suggestion to be a candidate when it appears at least in two of the three resources.

6 Conclusion and Future Work

We propose to integrate interactive machine learning for an annotation task and semantic writing aid application to incrementally train a classifier based on user feedback and interactions. While the goal of the annotation task is to produce a quality an-

_____
[4] http://goo.gl/C0YkiA

notated data, the classifier is built into the semantic writing aid application to continuously improve the system. The proposal addresses the following main points: 1) How to develop a quality linguistic dataset using interactive machine learning approach for a given annotation task. 2) How to systematically combine different NLP resources to generate paraphrase suggestions for a semantic writing aid application. Moreover, how to produce an application specific NLP resource iteratively using an interactive machine learning approach. 3) How to integrate user interaction and feedback to improve the effectiveness and quality of the system.

We have carried out preliminary experiments for creating sequence tagging data for German NER and Amharic POS. Results indicate that integrating interactive machine learning into the annotation tool can substantially reduce the annotation time required for creating a high-quality dataset.

Experiments have been conducted for the systematic integrations of different NLP resources (WordNet, PPDB, and JoBimText DT) as a paraphrasing component into a semantic writing aid application. Evaluation with the recall metric shows that the combination of resources yields better performance than any of the single resources.

For further work within the scope of this thesis, we plan the following:

- Integrate an active learning approach for the linguistic dataset development

- Investigate crowdsourcing techniques for interactive machine learning applications.

- Integrate more NLP resources for the semantic writing aid application.

- Investigate different paradigms of interactions, such as when and how the interactive classifier should produces new model and study how suggestions are better provided to annotators.

- Investigate how user interaction and feedback can improve the linguistic dataset development and the semantic writing aid applications.

- Investigate how to improve the paraphrasing performance by exploring machine learning for learning resource combinations, as well as by leveraging user interaction and feedback.

References

Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. Effective end-user interaction with machine learning. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA, 2011.

Darina Benikova, Chris Biemann, and Marc Reznicek. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

Rahul Bhagat and Eduard Hovy. What is a paraphrase? In *Association for Computational Linguistics*. MIT Press, 2013.

Chris Biemann. *Structure Discovery in Natural Language*. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-25922-7.

Chris Biemann and Martin Riedl. Text: now in 2D! A framework for lexical expansion with contextual similarity. *J. Language Modelling*, pages 55–95, 2013.

Igor A. Bolshakov and Alexander Gelbukh. Synonymous paraphrasing using wordnet and internet. In Farid Meziane and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, volume 3136 of *Lecture Notes in Computer Science*, pages 312–323. Springer Berlin Heidelberg, 2004.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. Parametric: An automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 97–104, Manchester, UK, 2008. Coling 2008 Organizing Committee.

Joana Costa, Catarina Silva, Mário Antunes, and Bernardete Ribeiro. On using crowdsourcing and active learning to improve classification performance. In *Intelligent Systems Design and Applications (ISDA)*, pages 469–474, San Diego, USA, 2011.

Koby Crammer and Yoram Singer. Ultraconservative online algorithms for multiclass problems. *J. Mach. Learn. Res.*, pages 951–991, 2003.

Christian Desrosiers and George Karypis. A comprehensive survey of neighborhood-based recommendation methods. *Recommender Systems Handbook*, 2011.

William Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. International Conference on Computational Linguistics, 2004.

Juri Ganitkevitch and Chris Callison-Burch. The multilingual paraphrase database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 4276–4283, 2014.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590, 2012.

Ali Ibrahim, Boris Katz, and Jimmy Lin. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 57–64, 2003.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. Enhancing clinical concept extraction with distributional semantics. In *Journal of Biomedical Informatics*, pages 129–140, San Diego, USA, 2012.

Kshitij Judah, Thomas Dietterich, Alan Fern, Jed Irvine, Michael Slater, Prasad Tadepalli, Melinda Gervasio, Christopher Ellwood, William Jarrold, Oliver Brdiczka, and Jim Blythe. User initiated learning for adaptive interfaces. In *IJCAI Workshop on Intelligence and Interaction*, Pasadena, CA, USA, 2009.

Mayank Kabra, Alice A Robie, Marta Rivera-Alba, Steven Branson, and Kristin Branson. Jaaba: interactive machine learning for automatic annotation of animal behavior. In *Nature Methods*, pages 64–67, 2013.

Anastasia Krithara, Cyril Goutte, MR Amini, and Jean-Michel Renders. Reducing the annotation burden in text classification. In *Proceedings of the 1st International Conference on Multidisciplinary Information Sciences and Technologies (InSciT 2006)*, Merida, Spain, 2006.

Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. Structured labeling to facilitate concept evolution in machine learning. In *Proceedings of CHI 2014*, Toronto, ON, Canada, 2014. ACM Press.

Alberto Lavelli, Bernardo Magnini, and Fabrizio Sebastiani. Building thematic lexical resources by bootstrapping and machine learning. In *Proc. of the workshop "Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data", wokshop at LREC-2002*, 2002.

Dekang Lin and Patrick Pantel. Dirt - discovery of inference rules from text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*, pages 323–328, San Francisco, CA, USA, 2001.

Barzan Mozafariy, Purnamrita Sarkarz, Michael Franklinz, Michael Jordanz, and Samuel Madden. Active learning for crowdsourced databases. In *arXiv:1209.3686*. arXiv.org preprint, 2012.

Daniel Naber. A rule-based style and grammar checker. diploma thesis, Computer Science - Applied, University of Bielefeld, 2004.

Roberto Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, pages 10:1–10:69, 2009. ISSN 0360-0300.

Ian Niles and Adam Pease. Toward a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9, 2001.

James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning*. O'Reilly Media, 2012. ISBN 978-1-4493-0666-3.

Hema Raghavan, Omid Madani, Rosie Jones, and Pack Kaelbling. Active learning with feedback on both features and instances. *Journal of Machine Learning Research*, 7, 2006.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin–Madison, 2010.

Burr Settles, Mark Craven, and Lewis Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008.

Ravi Sinha and Rada Mihalcea. Combining lexical resources for contextual synonym expansion. In *Proceedings of the International Conference RANLP-2009*, pages 404–410, Borovets, Bulgaria, 2009. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, 2008.

Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, pages 82–91, 2007.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, 2007. ISBN 978-1-59593-654-7.

Vedrana Vidulin, Marko Bohanec, and Matjaž Gams. Combining human analysis and machine data mining to obtain credible data relations. *Information Sciences*, 288:254–278, 2014.

Rui Wang and Chris Callison-Burch. Paraphrase fragment extraction from monolingual comparable corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 52–60, 2011.

Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H Witten. Interactive machine learning: letting users build classifiers. In *International Journal of Human-Computer Studies*, pages 281–292, 2002.

Julie Weeds and David Weir. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. In *Association for Computational Linguistics*, volume 31, pages 439–475, 2005.

Yi-Min Wen and Bao-Liang Lu. Incremental learning of support vector machines by classifier combining. In *Advances in Knowledge Discovery and Data Mining*, pages 904–911, Heidelberg, Germany, 2007.

Seid Muhie Yimam, Richard Eckart de Castilho, Iryna Gurevych, and Chris Biemann. Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 91–96, Stroudsburg, PA 18360, USA, 2014. Association for Computational Linguistics.