Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5678–5687
Marseille, 11–16 May 2020

A Large Harvested Corpus of Location Metonymy

Kevin Alex Mathews and Michael Strube
Heidelberg Institute for Theoretical Studies
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
{kevin.mathews, michael.strube}@h-its.org

Abstract

Metonymy is a figure of speech in which an entity is referred to by another related entity. The existing datasets of metonymy are either too small in size or lack sufficient coverage. We propose a new, labelled, high-quality corpus of location metonymy called WIMCOR, which is large in size and has high coverage. The corpus is harvested semi-automatically from the English Wikipedia. We use different labels of varying granularity to annotate the corpus. The corpus can directly be used for training and evaluating automatic metonymy resolution systems. We construct benchmarks for metonymy resolution, and evaluate baseline methods using the new corpus.
Keywords: Metonymy, Wikipedia disambiguation pages, DBpedia

1. Introduction

Metonymy is a figure of speech in which an entity is referred to by another related entity (Lakoff and Johnson, 1980; Littlemore, 2015). Consider the following two paragraphs:

(1) Nootdorp is a town in the Dutch province of South Holland. It is located approximately 2 km to the east of *Delft* and about 6 km southeast of the centre of The Hague.

(2) Against the wishes of his father, Hoff chose to study chemistry. First, he enrolled at *Delft* in September 1869, and studied until 1871 [...]

Note that the term *Delft* is used differently in the two paragraphs. In the former, *Delft* refers to the city of Delft in the Netherlands. In the latter, the same term does not refer to the city of Delft. From the context, it can be inferred that the term refers to an educational institution. In fact, it refers to the Delft University of Technology located in the city of Delft.

The linguistic phenomenon in action here is metonymy. Although the term *Delft* refers to the city in its literal sense, the same term refers to the university, a different but related entity, in its metonymic sense. While metonymy can operate on various kinds of names such as names of locations, organizations or persons, in this paper, we focus on location names only.

Metonymy is frequent in verbal as well as written communication. According to Gritta et al. (2017), about 20% of location names in the data sampled from Wikipedia are metonymically used. As a result, resolving metonymy aids various natural language processing (NLP) tasks such as machine translation (Kamei and Wakao, 1992), question answering (Stallard, 1993), named entity disambiguation (Harabagiu, 1998; Gritta et al., 2017), and coreference resolution (Fass, 1991). Further, metonymy is a universal phenomenon, and computational research has been conducted on data in different languages (Leveling and Hartrumpf, 2006; Poibeau, 2006).

The two existing datasets[1] on location metonymy are SEMEVAL (Markert and Nissim, 2007) and RELOCAR (Gritta et al., 2017). These corpora are small in size, containing about 2000 samples only. As a result, the datasets do not sufficiently cover the various ways in which metonymy can be observed in real-world data. Hence, these datasets are inadequate to be used for large-scale machine learning and statistical analyses. In addition, the samples in these datasets lack sufficient label granularity. For instance, the location *Delft* can be labelled as a place, and more specifically, as a city. This is what we mean by label granularity. As a result, there is disagreement over the annotations in the existing datasets.

WIKIPEDIA

Delft (disambiguation)

Delft is a city in the Netherlands.

Delft may also refer to **other places**:

- Delft, Cape Town, township in South Africa
- Neduntheevu, island in Sri Lanka
  - Delft Island fort, in Sri Lanka
  - Battle of Delft—a battle during the Sri Lankan Civil War
- Delft University of Technology, Dutch public university
- Delft, Minnesota, United States
- Delft Colony, California, United States

Material goods:

- Delft jewelry

Figure 1: Wikipedia disambiguation page for the topic *Delft*.

We harvested a new corpus called WIMCOR (Wikipedia Metonymy Corpus) using the English Wikipedia and DBpedia. We primarily employ the Wikipedia disambiguation pages to identify instances of metonymy. DBpedia is used to check the category of a Wikipedia entity. Finally, we generate sentences using these metonymic instances. Our corpus construction mechanism is semi-automatic in nature, with minimal human intervention. WIMCOR is an improvement over the existing datasets on various aspects such as size and label granularity. It is a testament to the richness and variety of metonymy. Every instance in the

[1] The terms dataset and corpus are used interchangeably in this paper.

new corpus is linked to a Wikipedia article and hence it alleviates any ambiguity over annotations, which is a drawback of the existing corpora. The new corpus can directly be used for training and evaluating automatic metonymy resolution systems.

The main contributions of this paper include the following: (1) present a new harvested corpus of location metonymy, (2) evaluate the corpus, and compare the corpus with the existing datasets, and finally (3) develop benchmarks for the task of metonymy resolution using the new corpus.

2. Related Work

In this section, we describe the key works that use Wikipedia as a resource. We also introduce the task of metonymy resolution and the existing datasets on metonymy.

2.1. Wikipedia

Wikipedia is a crowd-sourced, encyclopedic resource and is massive in size (as of 4 November 2019, there are over 5.9 million articles in the English Wikipedia). Wikipedia follows a semi-structured format through its use of infobox templates, table of contents inside articles, category network and disambiguation pages.

Wikipedia is used in NLP research because it is an excellent source of world knowledge. Gabrilovich and Markovitch (2007) computes word embeddings as weighted vectors of Wikipedia articles. Nastase et al. (2010) computes values of selectional preference features using the Wikipedia category network.

Wikipedia is also used to construct datasets for various tasks such as coreference resolution (Ghaddar and Langlais, 2016), conflict-of-interest detection (Orizu and He, 2018), and concept relatedness (Dor et al., 2018). Ghaddar and Langlais (2018) and Mihalcea (2007) exploit the internal hyperlink structure of Wikipedia to build large, annotated corpora for the tasks of fine-grained entity typing and word sense disambiguation respectively. Ge et al. (2018) creates a resource composed of the major events in human history using event-related infobox templates of Wikipedia. The structural information from articles is used to harvest inter-event relations.

2.2. Metonymy Resolution

The task of metonymy resolution aims to identify words that are used metonymically and interpret them appropriately. Markert and Nissim (2002) resolves metonymy using co-occurrences, collocations and grammatical features. Nastase and Strube (2009) and Nastase et al. (2012) use selectional preference features, which are computed using external resources such as British National Corpus, Wikipedia, WordNet (Miller, 1995) and WikiNet (Nastase et al., 2010). Gritta et al. (2017) proposes a neural-network-based model. This model is trained on a predicate window of context words, which is a set of words chosen with the the dependency head of the potentially metonymic word (PMW) as the starting point. The intuition behind the predicate window is the observation that the immediate context words are frequently noisy and hence, it is necessary to identify the right set of context words.

The two existing datasets on metonymy are SEMEVAL (Markert and Nissim, 2007) and RELOCAR (Gritta et al., 2017). The SEMEVAL data was sampled from BNC Version 1.0, and the RELOCAR (Real Location Retriever) data was sampled from Wikipedia. Both these corpora were compiled and labelled manually.

3. Corpus Details

In this section, we describe how we extract data to construct WIMCOR. We then present some basic details of the WIMCOR corpus.

In the rest of the paper, following the conventions in the literature, we refer to the word to be resolved as the PMW. Also, we refer to the title of a Wikipedia disambiguation page as the anchor text.

3.1. Resources

In this subsection, we briefly describe the two resources we use in our approach.

3.1.1. Wikipedia Disambiguation Pages

The Wikipedia disambiguation pages list different senses of ambiguous entities and provide links to Wikipedia articles corresponding to each sense of the entity. Consider the disambiguation page for the topic *Delft*, shown in Figure 1. The same term can refer to different entities such as a city in the Netherlands, a township in Cape Town, South Africa, or a city in Minnesota, United States. The English version of Wikipedia contains about 304,000 disambiguation pages. Mihalcea (2007) recognize the suitability of disambiguation pages as a sense inventory for the task of word sense disambiguation. In this work, we use the disambiguation pages to harvest terms that are used metonymically.

3.1.2. DBpedia Categories

DBpedia (Auer et al., 2007; Lehmann et al., 2015) is a large-scale knowledge base consisting of content extracted from Wikipedia. The English version of DBpedia describes 4.58 million *things*. We use DBpedia to check the category of a Wikipedia article. The DBpedia categorization is more structured and less noisy than Wikipedia's own category network. DBpedia is interlinked with various other resources such as YAGO (Suchanek et al., 2007) and Wikidata (Vrandečić and Krötzsch, 2014).

DBpedia assigns various categories to the entities in its knowledge base. The categories we use in our approach are listed here: (1) YAGO:LOCATION100027167 for locations such as towns, cities and countries, (2) WIKIDATA:Q3918 for educational institutions such as universities, (3) WIKIDATA:Q43229 for events such as battles and festivals, (4) WIKIDATA:Q4766028 for association football (soccer) teams and clubs, and (5) YAGO:STRUCTURE104341686 and YAGO:FACILITY103315023 for artifacts such as cathedrals, palaces and hospitals.

3.2. Corpus Construction

Figure 2 illustrates the process of corpus construction. The two main steps are extraction of metonymic pairs and generation of samples. The next two subsection describe each step in detail.

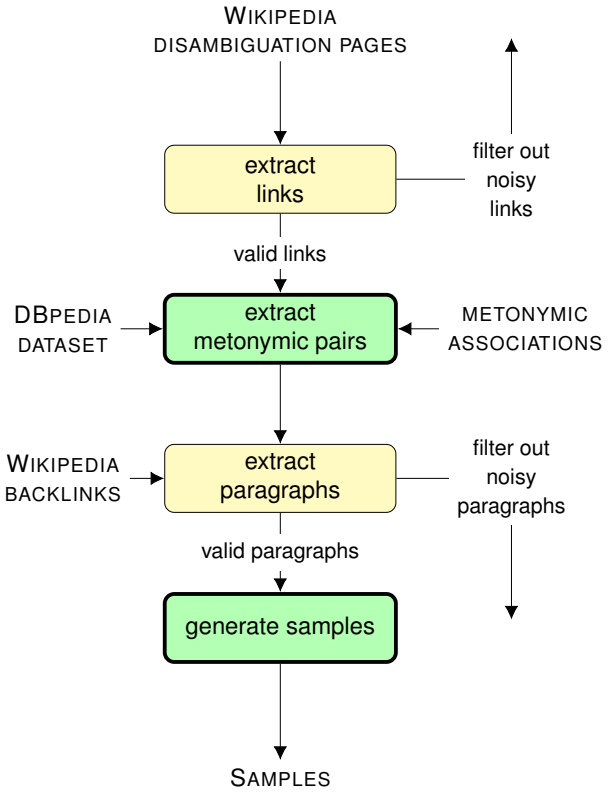

Figure 2: Flowchart of our corpus construction approach

Metonymic association	Metonymic pair	Anchor text
LOCATION-for-INSTITUTION	(*Delft*, *Delft University of Technology*)	Delft
LOCATION-for-TEAM	(*Milan*, *A.C. Milan*)	Milan
LOCATION-for-ARTIFACT	(*Arecibo*, *Arecibo Observatory*	Arecibo
LOCATION-for-EVENT	(*Busan*, *Busan International Film Festival*)	Busan

Table 1: The list of metonymic associations we use to extract data. A metonymic pair and the corresponding anchor text is also given.

3.2.1. Metonymic Pair Extraction

We use the Wikipedia disambiguation pages to harvest metonymic pairs. A metonymic pair $\langle W_L, W_M \rangle$ is a pair of Wikipedia articles that are referred to by the same *natural* title but denotes two different but strongly related concepts, such as Delft and Delft University of Technology. The samples in WIMCOR are generated using these metonymic pairs.

Note that metonymy is different from other linguistic phenomena such as homonymy or polysemy (Yarowsky, 1995). For instance, the city of Paris in France and Paris Hilton, the singer, do not form a metonymic pair because of the lack of any *strong* relationship between these two entities, although both entities can be referred to by the same term *Paris*. On the other hand, Delft and Delft University of Technology form a metonymic pair because the university is located in the city and both the city and the university can be referred to by the same term *Delft*.

As shown above, it is important to distinguish metonymic pairs in Wikipedia disambiguation pages from non-metonymic (polysemous) pairs. Our two-step method achieves this as follows: First, DBpedia is used to retrieve the category of a Wikipedia article and then check whether the pair matches a metonymic association. Table 1 presents a list of metonymic associations, along with an example pair and the anchor text (that is, the title of the corresponding Wikipedia disambiguation page). These associations are developed from commonly observed patterns of metonymy usage, and hence are useful means to organize instances of metonymy (Lakoff and Johnson, 1980; Radden and Kövecses, 1999). Secondly, we check whether the two articles refer to each other through internal hyperlinks. This is a simple but effective heuristic to ensure the existence of a *strong* relationship between two articles. The key intuition here is that strongly related Wikipedia articles tend to mention each other because of the encyclopedic nature of the resource. In this manner, we extracted hundreds of metonymic pairs for each association.

We queried Wikipedia offline using the XML dumps that were generated as of September 1st, 2019 and online using the MediaWiki API. We used the public SPARQL endpoint to query DBpedia remotely.

3.2.2. Sample Generation

After the metonymic pairs are extracted as explained above, we again use Wikipedia to generate samples. Note that any appropriate resource can be used for this purpose since the goal here is to extract titles of Wikipedia articles in context. Consider a 3-tuple (W_L, W_M, a_{lm}), where W_L and W_M denote any two Wikipedia articles that form a metonymic pair and a_{lm} denotes the corresponding anchor text. We generate samples $S(W_L, a_{lm})$ and $S(W_M, a_{lm})$ using the sample generator S, which is defined as follows:

$$S(W, a) = \{\, [w \mapsto a]\mathcal{P} \mid \mathcal{P} \in f(W) \,\}$$

where $f(W)$ traverses backlinks of the Wikipedia article W and extracts paragraphs having the title of the article W. The backlink of a Wikipedia article W points to other articles that contain mentions to W and has internal hyperlinks to it. The operation $[w \mapsto a]\mathcal{P}$ substitutes the hyperlinked mention w in the paragraph $\mathcal{P}$ with the anchor text a. Since a hyperlink to an article appears exactly once, the substitution operation applies to only a single mention in the paragraph. Figure 3 illustrates sample generation for the instance $S(\text{Delft University of Technology}, \textit{Delft})$.

In this way, from each tuple, it is possible to generate hundreds of samples. Up to 5000 backlinks can be retrieved using the MediaWiki API. We use some heuristics to filter out less useful and noisy paragraphs. For example, we ignore paragraphs that are either too short or too long by restricting the number of tokens in a sample to be between 10 and 512.

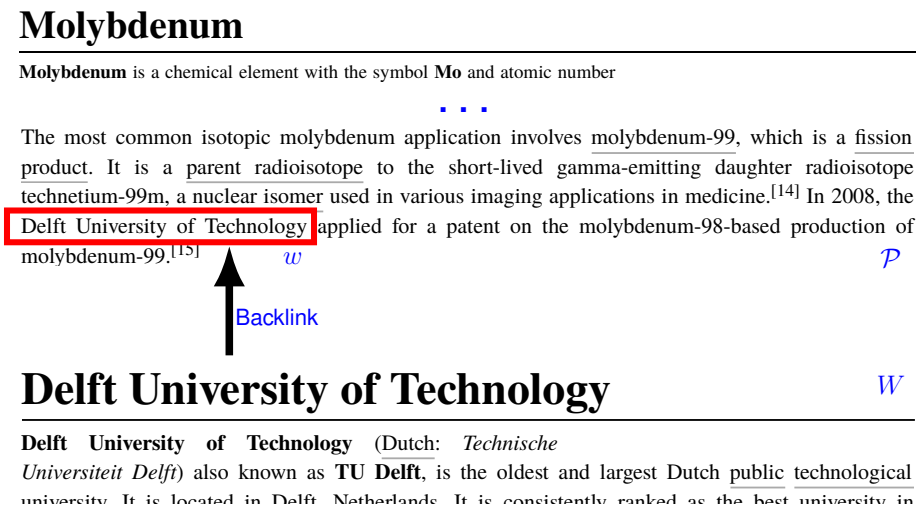

Figure 3: We extract paragraph $\mathcal{P}$ having mention w from the article *Delft University of Technology* (W) through its backlink. After replacing w with the anchor text, the following sample is generated: "The most common [. . .] medicine. In 2008, *Delft* applied for a patent on the molybdenum-98-based production of molybdenum-99."

3.3. Corpus Sample

The set of samples in WIMCOR is defined by $\forall_{(W_L,W_M,a_{lm})\in T}\,(\,S(W_L,a_{lm})\cup S(W_M,a_{lm})\,)$, where T is the list of tuples. Each sample, in general, is composed of more than one sentence. The PMW is marked explicitly in each sample. In addition, there are three labels of varying granularity: coarse-grained, medium-grained and fine-grained. All these labels can be used for classification. Some of the samples extracted by our approach are shown in Table 3.

The coarse-grained label identifies whether the PMW is literal or metonymic. The candidate labels in this case are LITERAL and METONYMIC. The samples in WIMCOR pertain to location names. As a result, if the PMW is interpreted to be a geographical entity, then it is labelled LITERAL. In the case of any other interpretation, the PMW is labelled METONYMIC. For example, the literal reading of the token *Delft* comprises the geographical and locative interpretations of the town Delft in the Netherlands. If the same token is used to denote the university located in the town Delft then the token is used metonymically. The medium-grained label identifies the entity type that the PMW refers to. The candidate labels in this case are entity types such as LOCATION, INSTITUTION, ARTIFACT, TEAM and EVENT. The fine-grained label identifies the specific entity that the PMW refers to. The candidate labels in this case are Wikipedia articles. The label hierarchy is shown in Figure 4.

3.4. Corpus Statistics

The raw corpus consists of more than 327K samples. In order to reduce noise, we retain only the samples corresponding to the pairs from which at least 50 samples were generated. The final version of the corpus is made up of 206K samples. The detailed corpus statistics are presented in Table 2.

In order to use the corpus with machine learning systems, we partition the corpus into train, validation and test in the ratio 60:20:20 respectively. For both SEMEVAL and RELOCAR, the data is separated in the ratio 50:50 for train and test respectively; there is no explicit validation set.

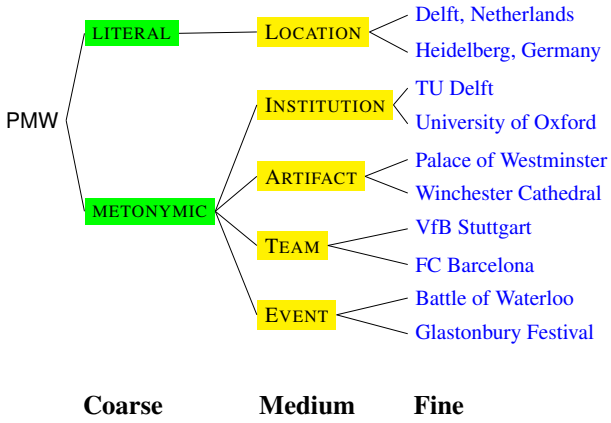

Figure 4: Label hierarchy for location names. The fine-grained labels shown here are by no means exhaustive.

The WIMCOR data and the code used for data extraction are available online[2]. The data is made available in three different formats: (1) a data file in XML format, (2) tabular data in CSV format, and (3) a dictionary in JSON format. The code is released under the GNU General Public License (version 3). The data is released under the Creative Commons Attribution-ShareAlike 3.0 Unported License.

4. Corpus Evaluation

In this section, we describe how we manually evaluate the harvested data. We then compare the corpus with the existing corpora of location metonymy.

4.1. Manual Evaluation

The WIMCOR corpus is generated and labelled automatically. So it is important to estimate the amount of noise in

[2] https://github.com/nlpAThits/WiMCor

	LOCATION-for-INSTITUTION		LOCATION-for-ARTIFACT		LOCATION-for-TEAM		LOCATION-for-EVENT		Total
Pairs	654		4023		639		88		5404
	LITERAL	METONYMIC	LITERAL	METONYMIC	LITERAL	METONYMIC	LITERAL	METONYMIC	
Raw Samples	63995	46941	94578	22802	71308	18324	6214	3269	327431
Min-50 Samples	52387	38332	72634	12068	59312	12980	4460	2175	254348
WIMCOR Samples	50000	30000	50000	10000	50000	10000	4000	2000	206000

Table 2: Corpus statistics, with respect to each metonymic association. Our approach extracts 5404 metonymic pairs and generates more than 327K samples. The final corpus is made up of 206K samples.

Text	She met Rich Annetts at the Glastonbury Festival in 2005. The couple moved to **Bath**, and lived in a flat close to the Royal Crescent.
Coarse	LITERAL
Medium	LOCATION
Fine	Bath, Somerset
Text	Wright taught astronomy and mathematics at **Elmira** before she was hired to be a computer at Harvard College Observatory.
Coarse	METONYMIC
Medium	INSTITUTION
Fine	Elmira College
Text	Radar results from **Arecibo** indicated that the comet nucleus was about 4.8 km (3.0 mi) across, and surrounded by a flurry of pebble-sized particles ejected at a few metres per second.
Coarse	METONYMIC
Medium	ARTIFACT
Fine	Arecibo Observatory
Text	Arsenal set a Champions League record during the 2005–06 season by going ten matches without conceding a goal, beating the previous best of seven set by **Milan**.
Coarse	METONYMIC
Medium	TEAM
Fine	A.C. Milan
Text	In 2012 Basu's film Barfi!, starring Ranbir Kapoor, Priyanka Chopra and Ileana D'Cruz, opened to largely-positive reviews and was well received at **Busan**.
Coarse	METONYMIC
Medium	EVENT
Fine	Busan International Film Festival

Table 3: Some samples of the WIMCOR corpus. The PMW is marked in boldface. Each sample is automatically assigned three labels of varying granularity, depending on the interpretation of the PMW in the given context.

the corpus and the reliability of the automatically assigned labels. For this purpose, we randomly selected 200 samples from the corpus. Two students of computational linguistics independently went through all these samples and evaluated the labels. Specific guidelines on the exercise were given to both the reviewers. The only acceptable response for each sample was either *Right* (that is, all the labels are correct) or *Wrong* (that is, at least one of the labels is incorrect). The responses received from the reviewers are presented in Table 4. In 88.50% of cases, both the reviewers agree that the automatically assigned labels are correct.

Despite having high actual agreement between the reviewers, the Cohen's kappa score (Cohen, 1960) is only 0.05, indicating very poor agreement. This is because the chance agreement between the reviewers is also very high, according to the assumptions made to compute Cohen's kappa. Byrt et al. (1993) proposes prevalence-adjusted bias-adjusted kappa (PABAK) score, to overcome some of the limitations of Cohen's kappa (Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990; Artstein and Poesio, 2008). The PABAK score of 0.78, in our case, indicates a strong agreement between the reviewers.

Coarse	Medium	Size	RR	WW	W
LITERAL	LOCATION	146	129	0	17
METONYMIC	INSTITUTION	29	26	0	3
	ARTIFACT	13	10	1	2
	TEAM	8	8	0	0
	EVENT	4	4	0	0
Total		200	177	1	22

Table 4: Responses received from the two label reviewers for a set of 200 samples selected randomly. RR denotes both the responses were *Right*. WW denotes both the responses were *Wrong*. W denotes at least one of the responses was *Wrong*.

4.2. Comparison with Existing Corpora

In this subsection, we compare WIMCOR with the existing corpora of location metonymy, namely SEMEVAL and RELOCAR. All the statistics reported in this section for all the corpora, unless otherwise stated, correspond to the train partition only.

Corpus	Size	Average Sample Length	Label Granularity
SEMEVAL	925	34	Coarse, Medium
RELOCAR	1026	22	Coarse
WIMCOR	124K	80	Coarse, Medium, Fine

Table 5: Comparison of WIMCOR with the existing corpora RELOCAR and SEMEVAL. WIMCOR is at least three orders of magnitude larger in size. In addition, the samples in WIMCOR are substantially longer. WIMCOR is annotated with three labels of varying granularity.

4.2.1. Quantitative Improvements

Table 5 compares WIMCOR with the existing corpora. WIMCOR is at least three orders of magnitude larger than both the existing corpora. The total number of samples in SEMEVAL and RELOCAR are 925 and 1026 respectively, while that in the WIMCOR corpus is 206K.

The average length of samples in WIMCOR is 80 tokens per instance. This is a major improvement when compared to SEMEVAL and RELOCAR, which have 34 and 22 tokens on average respectively. Figure 5 compares the length distributions of 500 samples randomly selected from WIMCOR, RELOCAR and SEMEVAL.

While the samples in SEMEVAL are annotated with coarse-grained and medium-grained labels (such as PERSON, EVENT, PERSON, FACILITY), the samples in RELOCAR are annotated with coarse-grained labels (LITERAL, METONYMIC and MIXED) only. In contrast, the samples in WIMCOR are annotated with coarse-grained, medium-grained and fine-grained labels. Table 6 compares the label distributions of samples in various corpora.

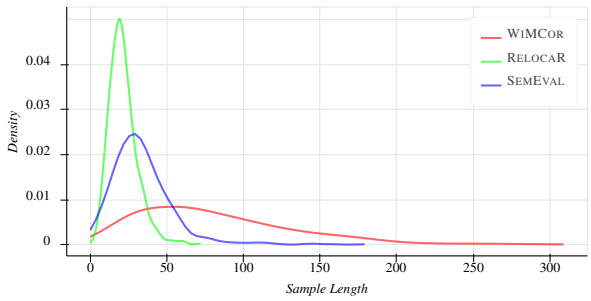

Figure 5: Comparison of length distributions of 500 samples randomly selected from WIMCOR, RELOCAR and SEMEVAL. The length of a sample is the number of tokens it contains. The samples in WIMCOR are substantially longer than those in RELOCAR and SEMEVAL.

4.2.2. Qualitative Improvements

Our corpus construction mechanism extracts a large number of unfamiliar and less popular metonymic pairs. This aspect can be attributed to the diversity of Wikipedia. Thus the WIMCOR corpus is a testament to the richness and variety of metonymy. For example, the majority of PMWs in the existing corpora are country names (82.05% and 73.88% for SEMEVAL and RELOCAR respectively). On the other hand, country names constitute only 10%, in the WIMCOR corpus. The rest of the PMWs is composed of a variety of location names such as names of towns (e.g. *Bath*), cities (e.g. *Freiburg*) and states (e.g. *Texas*). We used the DBpedia category YAGO:COUNTRY108544813 to check for country names.

Metonymy is a very difficult concept to grasp, even for experts. There is considerable disagreement over the existing annotations (Poibeau, 2007; Gritta et al., 2017). As a result, the annotation scheme of RELOCAR is different from that of the SEMEVAL data. While the former considers political entity interpretation of location names as metonymic, the latter considers them as literal. On the other hand, the fine-grained labels in the WIMCOR corpus are Wikipedia articles. Wikipedia articles provide encyclopedic information on a specific topic, and hence the target reading label alleviates the ambiguity over annotations to a large extend. We consider this feature of the WIMCOR corpus to be a major improvement over the existing corpora.

	Coarse	Medium	Count	%
SEMEVAL	LITERAL	LOCATION	737	79.7
	METONYMIC	PEOPLE	161	17.4
		EVENT	3	0.3
		PRODUCT	0	0.0
		MIXED	15	1.6
		OTHER	9	1.0
REL.	LITERAL		509	49.0
	METONYMIC		517	51.0
WIMCOR	LITERAL	LOCATION	154K	74.76
	METONYMIC	INSTITUTION	30K	14.56
		ARTIFACT	10K	4.85
		TEAM	10K	4.85
		EVENT	2K	0.97

Table 6: Comparison of coarse-grained and medium-grained label distributions of WIMCOR with that of the existing corpora SEMEVAL and RELOCAR. Note that the natural distribution of LITERAL and METONYMIC classes (in real-world data) in the case of location metonymy is approximately 80% and 20% respectively

5. Benchmarks for Metonymy Resolution

The WIMCOR corpus can be used to develop and evaluate metonymy resolution systems. Thus we create benchmarks for the task of metonymy resolution with the WIMCOR corpus.

In this paper, we address metonymy resolution as a multi-class classification problem. The target labels are the medium-grained labels: LOCATION, INSTITUTION, TEAM, ARTIFACT and EVENT. The objective, in this setting, is to identify the entity type referred to by the PMW in a given context.

5.1. Methods

We use various baseline methods such as uninformed, simple classifiers and informed classifiers that use different types of context to construct the benchmarks.

Method	Micro average			Macro average		
	Pre	Rec	F1	Pre	Rec	F1
Random	.587	.587	.587	.200	.200	.200
Majority	.749	.749	.749	.150	.200	.171
IMM5$_{GV}$	.870	.870	.870	.770	.508	.530
IMM10$_{GV}$	.890	.890	.890	.768	.550	.564
IMM50$_{GV}$	.900	.900	.900	.732	.578	.592
PREWIN$_{GV}$	.870	.870	.870	.746	.506	.522
IMM5$_{BT}$	.950	.950	.950	.900	.837	.860
IMM10$_{BT}$	.953	.953	.953	897	.837	.860
IMM50$_{BT}$	.950	.950	.950	.900	.823	.857
PREWIN$_{BT}$	.950	.950	.950	.883	.807	.850

Table 7: Performances of various baseline methods on the WIMCOR corpus. IMM and PREWIN refer to the methods based on immediate context and predicate window respectively. GV and BT denote the GLOVE and the BERT embeddings respectively.

5.1.1. Uninformed classifiers

We use two uninformed classifiers: a random classifier and a majority class classifier. The random classifier randomly picks a label from a list of class labels. The majority class classifier picks the label having the largest number of observations in the training set. These are simple classifiers because they do not make use of the context or any other information to make a decision. We use these classifiers to evaluate the performance of better informed classifiers.

5.1.2. Immediate context

The words surrounding the PMW are very useful in detecting its metonymicity. Traditional machine-learning techniques in metonymy resolution relied on context-based features such as co-occurrences and collocations (Markert and Nissim, 2002; Nissim and Markert, 2003; Markert and Nissim, 2005). The IMM baseline is a neural-network-based model that resolves metonymy using the immediate context of the PMW. We created three variants of the IMM baseline: IMM5, IMM10 and IMM50. The length of the context is different in each IMM variant. For instance, IMM5 uses a context of length 5 words, from either side of the PMW.

5.1.3. Predicate window

PREWIN (Gritta et al., 2017) employs a neural-network-based model. The model consists of four input layers in parallel: two LSTM layers for the right and left context words, and two dense layers for the dependency labels of the right and left context. A dropout of 0.2 is used in each input layer for regularization. The representations from the input layers are merged through concatenation. This merged representation is used for classification through a dense layer and a softmax layer. Categorical cross-entropy loss and adagrad optimizer are used for training the model. A batch size of 16 samples is used in training and testing the model. Finally, the model is trained for a total of 5 epochs. The hyper-parameters follow the original implementation. This model produced the state-of-the-art results on RELOCAR and SEMEVAL datasets with minimal use of external resources.

Note that the model architecture of PREWIN is the same as that of IMM. The primary difference between the two is the choice of context words. PREWIN uses a predicate window of context words, which is a set of words originating from the dependency head of the PMW. The key intuition behind the predicate window is that the immediate context words are frequently noisy and redundant. The predicate window is a small and focused set of context words. The length of the predicate window is set to 5.

5.2. Results

We evaluate each method using the following classification metrics: precision, recall and F1-score. The results for different word embeddings such as GLOVE (Pennington et al., 2014) and BERT (Devlin et al., 2019) are reported. In the case of GLOVE, the vocabulary is made up of only the top 100,000 most-frequent words. The zero vector is used to represent the out-of-vocabulary words. We use the pre-trained 50d (6B version) words vectors in our experiments. In contrast to GLOVE, BERT embeddings are context-sensitive and hence are able to distinguish, for example, different senses of polysemous words from each other. We use the pre-trained base, uncased version of BERT in our experiments. Instead of deploying BERT as a classifier, we use the PREWIN model and initialize it with the GLOVE-like word embeddings using BERT. For this purpose, we concatenate the representations from the last four hidden layers of the BERT transformer to compute subword embeddings. Note that the BERT tokenizer splits a word into one or more subwords. These subword embeddings are then combined through summation to generate GLOVE-like word embeddings. There are no out-of-vocabulary words in the case of BERT.

The results are presented in Table 7. Since there is class imbalance in the data, we report the micro-averaged and macro-averaged metrics. The IMM and PREWIN baselines outperform the uninformed classifiers. In addition, BERT is better than GLOVE because of the context-sensitive nature of the BERT embeddings. For both GLOVE and BERT, the larger the context, the better the performance.

The high performance on WIMCOR, especially using the BERT embeddings, does not mean that the dataset is trivial or easy to solve. There are two main reasons for this conclusion. First, while macro-averaging treats all classes alike, micro-averaging takes into account the proportion of each class in the data. The classifiers perform well for the majority class LOCATION, but do not exhibit a similar performance for the other classes. As a result, the macro-averaged results of all the methods are low when compared to the corresponding micro-averaged results. Second, the PMW is specified explicitly in the current experimental setting. This setting is easier because the classifier can exploit the context words that are indicative of a particular class to classify the PMW, while not performing anything strictly relevant to metonymy resolution (Zellers et al., 2019). A more challenging setting is where the PMW is not specified in advance (Mao et al., 2019). WIMCOR can be made to fit this new setting by considering every word (or phrase) in a sample as a PMW. This setting also enables a metonymy resolution system to be deployed on real-world data and be used for downstream tasks.

5.3. Discussion

5.3.1. Limitations of our Approach

Metonymy is a linguistic phenomenon that can manifest itself in language in different ways such as multi-word expressions, proper nouns and common nouns (Littlemore, 2015). For instance, in the sentence "We need a couple of strong bodies on our team." (Lakoff and Johnson, 1980), the noun phrase *strong bodies* is used metonymically to mean people with strong bodies. However, our extraction mechanism identifies the metonymic pairs that are listed on the Wikipedia disambiguation pages only, which in turn are Wikipedia articles. While proper nouns of types LOCATION or PERSON have articles, common nouns do not have articles in Wikipedia. As a result, the PMWs in our corpus are limited to proper nouns only.

In our approach, errors creep in either from Wikipedia or DBpedia. For instance, according to Wikipedia, the term *Westlake* can refer to *Westlake Girls High School* or *Westlake Boys High School*. However according to DBpedia, *Westlake Girls High School* is categorized as a location. As a result, for the association LOCATION-for-INSTITUTION our extraction mechanism incorrectly extracts the following metonymic pair: *Westlake Girls High School* for *Westlake Boys High School*, which means the former can denote the latter by way of metonymy. So all the samples generated from this pair turn out to be false positives. Furthermore, the disambiguation pages and Wikipedia at large are primarily written keeping human users in mind. While extracting metonymic pairs, we have used various heuristics to filter out noisy articles. However, it is difficult to achieve 100% accuracy at this step.

5.3.2. Future of Metonymy Resolution Research

Metonymy resolution can be decomposed into two steps (Markert and Nissim, 2002): metonymy detection (check whether the PMW is literal or metonymic in the given context) and metonymy interpretation (identify the actual target entity, if the PMW is metonymically used). Existing metonymy resolution systems tend to focus on metonymy detection only, mainly because of the shortage of labelled data and the lack of sufficient label granularity in the existing data. WIMCOR comprises thousands of labelled samples. In addition, the fine-grained label in each sample identifies the target entity in terms of Wikipedia articles, which is as specific as it can be. This information aids research on metonymy interpretation.

6. Conclusions

In this paper, we introduce a new corpus of metonymy called WIMCOR. The corpus is generated semi-automatically using disambiguation pages of the English Wikipedia. The WIMCOR corpus is an improvement over the existing corpora on various aspects such as size and label granularity. We publish benchmarks using the new corpus for the task of automatic metonymy resolution. The multilingual nature of Wikipedia offers the possibility to extend WIMCOR to many other languages. We believe the new corpus will aid the study on metonymy and automatic metonymy resolution systems.

7. Acknowledgements

We thank Fabian Düker and Jason Brockmeyer for assistance in manual evaluation. We would like to acknowledge Federico López for the helpful discussion and suggestions. We also thank the anonymous reviewers for their comments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a HITS PhD scholarship.

8. Bibliographical References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, pages 722–735. Springer-Verlag.

Byrt, T., Bishop, J., and Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423 – 429.

Cicchetti, D. V. and Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551 – 558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dor, L. E., Halfon, A., Kantor, Y., Levy, R., Mass, Y., Rinott, R., Shnarch, E., and Slonim, N. (2018). Semantic relatedness of wikipedia concepts – benchmark data and a working solution. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Fass, D. (1991). Met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543 – 549.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1606–1611. Morgan Kaufmann Publishers Inc.

Ge, T., Cui, L., Chang, B., Sui, Z., Wei, F., and Zhou, M. (2018). EventWiki: A knowledge base of major events. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Ghaddar, A. and Langlais, P. (2016). WikiCoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 136–142. European Language Resources Association (ELRA).

Ghaddar, A. and Langlais, P. (2018). Transforming Wikipedia into a large-scale fine-grained entity type corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2017). Vancouver welcomes you! minimalist location metonymy resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1248–1259. Association for Computational Linguistics.

Harabagiu, S. M. (1998). Deriving metonymic coercions from WordNet. In *Usage of WordNet in Natural Language Processing Systems*. Association for Computational Linguistics.

Kamei, S.-i. and Wakao, T. (1992). Metonymy: Reassessment, survey of acceptability, and its treatment in a machine translation system. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 309–311. Association for Computational Linguistics.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press, Chicago.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Leveling, J. and Hartrumpf, S. (2006). On metonymy recognition for geographic ir. In *Proceedings of GIR-2006: 3rd ACM Workshop on Geographical Information Retrieval*. Department of Geography, University of Zurich.

Littlemore, J. (2015). *Metonymy: Hidden Shortcuts in Language, Thought and Communication*. Cambridge University Press, Cambridge.

Mao, R., Lin, C., and Guerin, F. (2019). End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898. Association for Computational Linguistics.

Markert, K. and Nissim, M. (2002). Metonymy resolution as a classification task. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 204–213. Association for Computational Linguistics.

Markert, K. and Nissim, M. (2005). Learning to buy a Renault and talk to BMW: A supervised approach to conventional metonymy. In *Proceedings of the 6th International Workshop on Computational Semantics*.

Markert, K. and Nissim, M. (2007). Semeval-2007 task 08: Metonymy resolution at semeval-2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41. Association for Computational Linguistics.

Mihalcea, R. (2007). Using Wikipedia for automatic word sense disambiguation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 196–203. Association for Computational Linguistics.

Miller, G. A. (1995). WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nastase, V. and Strube, M. (2009). Combining collocations, lexical and encyclopedic knowledge for metonymy resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing - Volume 2*, pages 910–918. Association for Computational Linguistics.

Nastase, V., Strube, M., Börschinger, B., Zirn, C., and Elghafari, A. (2010). WikiNet: A very large scale multilingual concept network. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1015–1022. European Language Resources Association (ELRA).

Nastase, V., Judea, A., Markert, K., and Strube, M. (2012). Local and global context for supervised and unsupervised metonymy resolution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 183–193. Association for Computational Linguistics.

Nissim, M. and Markert, K. (2003). Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 56–63. Association for Computational Linguistics.

Orizu, U. and He, Y. (2018). Content-based conflict of interest detection on Wikipedia. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Poibeau, T. (2006). Dealing with metonymic readings of named entities. In *Proceedings of COGSCI'06*, pages 1962–1968. Cognitive Science Society.

Poibeau, T. (2007). UP13: Knowledge-poor methods (sometimes) perform poorly. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 418–421. Association for Computational Linguistics.

Radden, G. and Kövecses, Z. (1999). Towards a theory of metonymy. In *Metonymy in Language and Thought*, volume 4, pages 17–60. John Benjamins Publishing.

Stallard, D. (1993). Two kinds of metonymy. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 87–94. Association for Computational Linguistics.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. In *16th International World Wide Web Conference, WWW2007*, pages 697–706. ACM Press.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800. Association for Computational Linguistics.