*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 2617–2625
Marseille, 20-25 June 2022

# WiC-TSV-de: German Word-in-Context Target-Sense-Verification Dataset and Cross-Lingual Transfer Analysis

**Anna Breit, Artem Revenko, Narayani Blaschke**
Semantic Web Company
Neubaugasse 1, Vienna, Austria
{anna.breit, artem.revenko}@semantic-web.com

## Abstract

Target Sense Verification (TSV) describes the binary disambiguation task of deciding whether the intended sense of a target word in a context corresponds to a given target sense. In this paper, we introduce WiC-TSV-de, a multi-domain dataset for German Target Sense Verification. While the training and development sets consist of domain-independent instances only, the test set contains domain-bound subsets, originating from four different domains, being *Gastronomy*, *Medicine*, *Hunting*, and *Zoology*. The domain-bound subsets incorporate adversarial examples such as in-domain ambiguous target senses and context-mixing (i.e., using the target sense in an out-of-domain context) which contribute to the challenging nature of the presented dataset. WiC-TSV-de allows for the development of sense-inventory-independent disambiguation models that can generalise their knowledge for different domain settings. By combining it with the original English WiC-TSV benchmark, we performed monolingual and cross-lingual analysis, where the evaluated baseline models were not able to solve the dataset to a satisfying degree, leaving a big gap to human performance.
WiC-TSV-de data is openly available at `https://github.com/semantic-web-company/wic-tsv/`.

**Keywords:** Target Sense Verification, Disambiguation, German, Dataset, Cross-Lingual Disambiguation

## 1. Introduction

Being able to distinguish between different meanings of a word is a crucial pre-processing step for a wide variety of down-stream tasks such as document retrieval, sentiment analysis, or relation extraction.

Traditionally, the formulation of the *Word Sense Disambiguation (WSD)* task was used to tackle this problem, where for a given target word, the corresponding best-matching entry from an external word sense inventory was retrieved. Though this task formulation enjoys great popularity, it comes with downsides that especially show in domain-specific, real-world scenarios: since the target senses are linked to external inventories, WSD systems need to be modelled accordingly to these resources, which not only reduces the flexibility of the models, but also enforces the assumption of the availability of complete data within these resources. In certain scenarios however, it might be impractical or even impossible to model all possible senses.

The more recent task formulation of *Word-in-Context (WiC)* breaks with this dependencies on external sense inventories by focusing on the question whether two contexts use the target word in the same sense. This binary formulation increases the flexibility, but focuses on identifying similar usages of a target instead of assigning a specific sense to a target. In domain-specific and enterprise settings, the focus often lies on a defined set of concepts, i.e., a relatively small subset of all possible senses, which needs to be identified and disambiguated within textual data.

The task formulation of *Target Sense Verification (TSV)* combines the independence of sense inventories and the possibility to identify and assign specific senses, by formulating the disambiguation as a binary classification task: Given a context and a target sense, the task is to verify whether the target word in the context is used in the target sense.

In order to create well-performing and reliable TSV models, high-qualitative datasets are necessary, both for their training and evaluation process. However, the number of such datasets is still extremely low: to the best of our knowledge, the WiC-TSV dataset (Breit A. et. al., 2021) is the only resource created for this purpose, which is available in English only. As the creation of large-scale datasets in different languages forms a major bottleneck for the advancement of this area (Pasini, 2020), the development of systems that are capable of performing language transfer on the task are preferable, especially for low-resourced languages.

In this paper, we are presenting WiC-TSV-de, a German Word-in-Context Target-Sense-Verification dataset, which not only enables the training and analysis of German TSV models, but also –in combination with the English version– pathens the way for evaluating the cross-lingual transfer capabilities of these approaches. Due to domain-bound test subsets, this resource further allows the evaluation of how well a given model can adapt to a certain domain.

The rest of the paper is structured as follows. Section 2 gives an overview of existing disambiguation datasets, while Section 3 provides a formal introduction to the TSV task and German sense inventories. In Section 4, we elaborate on the creation process of WiC-TSV-de and its characteristics, followed by the analysis of mono-lingual and cross-lingual baseline models in Section 5 and 6. We conclude our findings in Section 7.

## 2. Related Work

**Disambiguation Datasets** There is a rich number of evaluation datasets targeting different aspects of the standard WSD task, e.g., broad, general frameworks (Raganato et al., 2017; Vial et al., 2018) domain-specific (Agirre et al., 2010; Faralli and Navigli, 2012) and language specific ones (Henrich and Hinrichs, 2012; Okumura et al., 2010; Scarlini et al., 2020).
In 2019, (Pilehvar and Camacho-Collados, 2019) added a new flavour to the disambiguation model evaluation landscape by introducing the Word-in-Context dataset, where the focus lies on deciding whether or not the same sense is used in two different contexts. WiC was also integrated as a subtask in the general language understanding framework SuperGLUE (Wang et al., 2019). Further extended WiC datasets were introduced by (Raganato et al., 2020), (Liu et al., 2021), and (Martelli et al., 2021), for more details, see below. Finally, (Breit A. et. al., 2021) introduced the initial English WiC-TSV dataset, which not only allows the evaluation of the generalisation capabilities of TSV-models, but also their ability to adapt to certain domains, as different domain-specific test subsets were included.

**Cross-Lingual Transfer Analysis** The analysis of cross-lingual transfer capabilities in disambiguation models has a long tradition –with according SemEval tasks already being published 15 years ago showing interest in this area (Agirre et al., 2007)– and has not lost its relevance today (Gella et al., 2019; Procopio et al., 2021; Ataman et al., 2021). This trend is also visible for recent WiC datasets. XL-WiC (Raganato et al., 2020) contains training sets in three different languages as well as evaluation sets for 12 different languages with varying degrees of resource availability. As a further extension, AM2iCo (Liu et al., 2021) provides datasets from 14 different languages, including difficult adversarial examples, as well as corresponding training sets for 10 of these languages, enabling diverse cross-lingual analysis. The same year, MCL-WiC (Martelli et al., 2021) has been introduced as an entirely manually-annotated dataset for multi- and cross-lingual WiC, whose evaluation sets are available in 5 different languages.

## 3. Preliminaries

**TSV** The task of Target Sense Verification can be formally described as the binary classification of an instance $x$, where each $x$ consists of a context $c$ containing a target word $w$, and a target sense $s_w$ represented by one or multiple sets of sense descriptors $d_s$. The classification aims at determining whether the intended sense of the word $w$ used in the context $c$ matches the target sense $s$. Sense descriptors could be of varying types such as descriptions, hypernyms, or synonyms. In a multilingual setting, the target sense could also be indicated by e.g., its translations.
For the described task, an English dataset was published in previous work (Breit et al., 2021) containing 3832 instances, where the training and development set consist of domain-independent instances only (retrieved from WordNet and Wiktionary), while the test set also includes three domain-specific subsets, being *Cocktails*, *Medicine*, and *Computer Science*. As sense descriptors, definitions and hypernyms collected from existing Semantic Web resources were provided. The data of WiC-TSV can be found at `https://github.com/semantic-web-company/wic-tsv`

**German Sense Inventories** For English disambiguation tasks, the most prominent resource is WordNet (Fellbaum, 1998), a large lexical database that collects words, associates them with meanings by grouping them in so-called synsets, and organises these sets by the means of conceptual-semantic and lexical relations. The idea of WordNets was adopted for many different languages, including German, where two major resources exist: GermaNet (Hamp and Feldweg, 1997) and OdeNet (Siegel and Bond, 2021).

*GermaNet* was initiated in 1996 by the Univeristy of Tübingen, Germany, and has since then grown to a mature online thesaurus, organising more than 150,000 synsets. However, the resource underlies rather strict licensing requirements, making the easy re-use, and distribution difficult.

*OdeNet*, on the other hand, is an initiative that aims at providing German thesauric knowledge in an open and easily accessible way. However, as this resource is still in developmental state, certain aspects are not fully covered yet: while OdeNet contains more than 120,000 lexical entries in about 36,000 synsets, the resource only contains approximately 19,000 definitions and not even 1000 examples.

A different approach of building a lexical resource is taken by *Wiktionary*, an online dictionary available in a wide variety of languages, where the content is collected in a collaborative, crowd-sourced way. In comparison to expert-built lexicons, Wiktionary is therefore more coarse-grained, as the entries focus more on the general understanding of meanings, than on the linguistic correctness. The collaborative approach on the one hand helps to keep the resource up-to-date with new terms and term usages, and on the other hand improves the coverage of domain specific word senses. Wiktionary not only contains words and their different meanings, but also provides templates to add synonyms, examples sentences, and hypernym and hyponym relations for these senses. However, since entries in this resource are represented by encyclopedic-style pages, the content of these entries is only provided in an semi-structured way, and is prone to errors. Still, with over 1.1M German entries, Wiktionary plays an important role as sense inventory.

| Tag | Context | Definition | Hypernyms |
|---|---|---|---|
| **Domain-independent (WIK)** | | | |
| T | Sie zücken ihre Handys und Adressbücher, bestehen darauf, dass ich die Namen und Adressen ihrer Verwandten in Europa in meinen **Block** schreibe.<br>They pull out their mobile phones and address books, insist that I write the names and addresses of their relatives in Europe in my **notepad**. | ein Stapel Papierblätter, welche miteinander verklebt oder verdrahtet sind und nach Bedarf abgerissen werden können.<br>a stack of sheets of paper which are glued or wired together and can be torn off as required. | Schreibware<br>stationery |
| F | Der **Ruf** “Feuer!” hallte durch das Haus .<br>The **call** “Fire!” echoed through the house. | das Ansehen, das jemand bei anderen hat<br>the reputation that someone has among others | Ansehen, Status<br>reputation, status |
| **Gastronomy (FOOD)** | | | |
| T | Betrachten wir nun ein Objekt, beispielsweise einen Apfel, treffen die von diesem **Apfel** reflektierten Lichtstrahlen auf unsere Hornhaut.<br>If we now look at an object, for example an apple, the light rays reflected from this **apple** hit our cornea. | Frucht des Apfelbaums<br>fruit of the apple tree | Kernobst<br>Pome fruit |
| F | Iris **Apfel** ist bekannt für ihren exzentrischen Stil, mit dem sie dem Jugendwahn seit Jahren den Spiegel vorhält.<br>Iris **Apfel** is known for her eccentric style, with which she has been holding up a mirror to the youth craze for years. | Frucht des Apfelbaums<br>fruit of the apple tree | Kernobst<br>Pome fruit |
| **Hunting (HUNT)** | | | |
| T | Bemerken möchte ich dazu, dass die Katzen und vorjährigen Murmel viel mehr schreien als die älteren **Bären**.<br>I would like to add to this that the female marmots and one-year-old marmots scream much more than the older **male marmots**. | Die Bezeichnung für das männliche Murmeltier<br>The name for the male marmot | Niederhaarwild<br>small wild game |
| F | Wie auch Wildkatze, **Bär** oder Wolf haben die Menschen den Luchs in der Vergangenheit intensiv gejagt und ihm den Lebensraum streitig gemacht.<br>Like the wildcat, **bear** or wolf, humans have intensively hunted the lynx in the past and dispossessed it of its habitat. | Die Bezeichnung für das männliche Murmeltier<br>The name for the male marmot | Niederhaarwild<br>small wild game |
| **Medicine (MED)** | | | |
| T | Metaphysenbrüche am linken Schienbein und an der rechten **Elle**.<br>Metaphyseal fractures of the left tibia and right **ulna**. | Kleinfingerseitig gelegener länglicher Röhrenknochen des Unterarms<br>Elongated tubular bone of the forearm on the small finger side | Armknochen<br>arm bone |
| F | Mit diesem Beitrag sollen die Fakten aufgezeigt werden, die das Einmessen mit der ägyptischen **Elle** bezeugen.<br>This article intends to show the facts documenting the calibration using the Egyptian **cubit** | Kleinfingerseitig gelegener länglicher Röhrenknochen des Unterarms<br>Elongated tubular bone of the forearm on the small finger side | Armknochen<br>arm bone |
| **Zoology (ZOO)** | | | |
| T | Aus einem Rappen wird ein Rappfalbe, beziehungsweise ein Graufalbe, aus einem Braunen ein Braunfalbe und aus einem **Fuchs** ein Rotfalbe.<br>A black horse becomes a black dun, or a grey dun, a brown horse becomes a brown dun and a **chestnut** becomes a red dun. | Fellfarbe. Rotbraunes Fell mit gleichfarbiger Mähne und Schweif.<br>Coat colour. Reddish brown coat with mane and tail of the same colour. | Pferde nach Fellfarbe<br>Horses by coat colour |
| F | Ein einzelnes Weibchen des Kleinen **Fuchses** legt nach der Überwinterung im März oder April etwa 150 Eier auf ein Brennnesselblatt.<br>A single female of the small **tortoiseshell** lays about 150 eggs on a nettle leaf after hibernation in March or April. | Fellfarbe. Rotbraunes Fell mit gleichfarbiger Mähne und Schweif.<br>Coat colour. Reddish brown coat with mane and tail of the same colour. | Pferde nach Fellfarbe<br>Horses by coat colour |

Table 1: Sample instances from the WiC-TSV-de dataset. Target words are marked in bold within the contexts. Tags: T (True) and F (False). Below each original instance, an English translation is provided.

## 4. WiC-TSV-de: The Data Set

Following the definition of the TSV task, we created a dataset of 4117 instances in German language. Each instance consists of a context containing the target word and two different kinds of target sense descriptors: hypernyms and definition (see Table 1). Target words are all single-word nouns, contexts and definitions are tokenised. While the training and the development set consist only of *domain-independent* instances, i.e., they do not focus on a certain domain, the test set also contains *domain-bound* subsets, where the target senses are domain-specific. The construction of these different kinds of instances as well as the characteristics of the dataset are described below.

## 4.1. Dataset Construction

### 4.1.1. Domain-Independent Instances

Domain-independent instances were constructed from the German Wiktionary. For this purpose, after an initial cleaning step, we scraped all entries of German nouns that have a definition and at least one example associated with it. Entries where only one sense fulfilled the above criteria were either rejected or used as positive example, while entries with multiple valid senses produced both positive and negative examples. Negative examples were created by randomly interchanging the target senses and examples. During the creation process, examples in which the target word did not appear (e.g., because a synonym was used), or for which the sense numbering was incorrect (i.e., the example was annotated to be connected to sense *[2]*, but there was no corresponding sense *[2]*) were removed. Great care was taken to avoid information leakage (e.g., created by using the same example with two different target senses) and to keep the number of positive and negative examples balanced. This procedure resulted in 3537 instances, that were split with a 72:12:16[1] ratio into training, development and test set.

### 4.1.2. Domain-Bound Instances

For the domain-bound instances, first a set of domain-specific ambiguous nouns and their domain-specific target senses was manually created. Then, example contexts were collected manually by incorporating search engines. The contexts which were retrieved from a variety of resources including news articles, blog posts, and recipes, were required to be written in proper German, whereby the formality of the language used was incidental. A further restriction regarding the resources constricted the collection from dictionary and encyclopedia content, in order to maintain a realistic test use case. Domain-bound instances were collected for the following domains:

**Gastronomy (FOOD)** For this domain, ambiguous names of (parts of) fruits and vegetables (32%), dishes (44%), meat cuts (20%), and tools (4%) were identified as target words. Due to the diversity of the field, we could not identify a single resource to retrieve information about definitions and hyperyms for all the targets, but used separate online resources for fruits, vegetables and dishes[2] [3], and tools and meat cuts[4] [5].

**Hunting (HUNT)** This subset is based on the vocabulary used by hunters to describe body parts or subcategories of game (76%), as well as their tracks (10%), and hunting methods (14%). While we collected the definitions from an online resource[6], the hypernyms were added manually, originating from the agreement of two domain experts.

| | | **Total** | $N_w$ | $R_+$ |
|---|---|---|---|---|
| **Train** | **WIK** | 2532 | 1989 | 0.51 |
| **Dev** | **WIK** | 425 | 405 | 0.49 |
| **Test** | **All** | 1160 | 633 | 0.50 |
| | **Domain-independent**<br>(WIK) | 580 | 548 | 0.50 |
| | **Domain-bound**<br>**(FOOD+HUNT+MED+ZOO)** | 580 | 91 | 0.50 |
| | **FOOD** | 145 | 25 | 0.50 |
| | **HUNT** | 140 | 21 | 0.50 |
| | **MED** | 140 | 22 | 0.50 |
| | **ZOO** | 155 | 25 | 0.51 |

Table 2: Statistics of training, development and testing splits of WiC-TSV-de, including total number of instances (**Total**), unique number of target words ($N_w$) and percentage of positive instances ($R_+$).

**Medicine (MED)** For the medical domain, targets describing body parts (73%) and illnesses (27%) were used. Definitions are derived from Pschyrembel[7], a common German medical dictionary, while hypernyms are taken from the German version of MeSH[8].

**Zoology (ZOO)** The zoology subset consists of ambiguous terms for animals (84%) and animal body parts (16%). While definitions were collected from the rominated reference book *Lexicon of Biology*[9](Sauermost and Freudig, 1998), the corresponding hypernyms where retrieved from the taxonomy provided by a German animal Encyclopedia[10].

## 4.2. Dataset Characteristics

### 4.2.1. Quantitative Characteristics

A statistical overview of the dataset and its splits is shown in Table 2. The totality of 4117 available instances were split into training, development and testing sets with a ratio of 62:10:28 which allows for a sophisticated analysis of model performance with respect to their generalisation capabilities, while still providing an appropriately sized training set.

While the training and the development sets are domain-independent, half of the test set consist of domain-specific examples with 140-155 instances per domain. For the domain-bound subsets, the average number of instances per target word is with 5-7 significantly higher than for the domain-independent set.

[1] Please note that these are the ratios of the domain-independent instances only, not for the entire dataset.

[2] https://www.faz.net/aktuell/stil/essen-trinken/

[3] https://www.dwds.de/wb/

[4] https://www.fleischwirtschaft.de/fachbegriffe

[5] https://www.gastroinfoportal.de/lexikon/

[6] https://www.jagdschulatlas.de/jagdlexikon/

[7] https://www.pschyrembel.de/

[8] https://www.dimdi.de/dynamic/de/klassifikationen/weitere-klassifikationen-und-standards/mesh/

[9] https://www.spektrum.de/lexikon/biologie/

[10] http://www.tierdoku.com/

Due to the manual creation process, it is ensured that the different instances for one target are of high variety and therefore redundancies in the test set are prevented despite the higher number of examples. For all splits, the number of positive and negative instances is approximately balanced.

While the target *word* overlap of the test and train set is about 36% of the instances, only 14% of target *senses* appearing in the test set are also used in the training set. Remarkably, there is no overlap between the domain-bound (test) target senses and the target senses in the training set. The overlap between the different test subsets is neglectable: seven domain-independent instances use target words, that are also present in one of the domain-bound sets, where only a single one of these instances targets the same domain-specific sense.

### 4.2.2. Qualitative Characteristics

Apart from the purely quantitative statistics, taking a closer look at the qualitative characteristics of the dataset and its subset can help interpreting the performance of tested models.

Even though examples taken from Wiktionary are not bound to a specific domain, they may include instances that are domain-specific due to the broad coverage of the resource. We did not filter these examples, as the presence of some domain-specific instances does not contradict the assumption that the instances are retrieved from a general domain corpus, since these corpora do not categorically exclude the usage of domain-specific terminology. However, great care was taken that the target senses included in the test set are not present in the training and development set.

Taking a look at the target senses included in the domain-bound test-subsets, we can see differences in their commonality: In *Medicine* for example, for most of the instances, domain-specific terms that are familiar to a broad, non-expert audience are used, as higher specialised terms are only ambiguous to a limited extent because German incorporates many foreign terms in their medical language. On the other hand, for the instances in the *Hunting* domain, a highly specialised terminology is used, where the target sense never corresponds to the most common sense of the target word. For *Gastronomy* and *Zoology*, part of the target senses (e.g., fruits and vegetables in *Gastronomy*) describe the most commonly used sense of the target word, while others are lesser known senses.

For quantifying this commonality we analysed for each domain-bound instance in the dataset (a) whether the target sense corresponds to the first sense of the target word listed in Wiktionary and (b) whether the target sense is listed at all in Wiktionary or in Duden[11] – a well established German dictionary. Here, we use the ordering of the senses in Wiktionary as an estimate for their commonality, being well-aware that due to the collaborative nature of this resource the equation of the first with the most frequent meaning is merely an assumption, and not enforced quality[12]. A summary of the target sense commonality for the different subsets can be found in Table 3.

In order to further create highly challenging test instances, different strategies were implemented when manually collecting contexts for the domain-bound target senses. Of special interest are target words that have a second meaning within the same (or a neighbouring) domain. For example, in the hunting domain the term “Bär” is both used to refer to a *bear* as well as to a *male marmot*. These ***in-domain ambiguities*** are a great starting point to generate highly challenging test instances, where it is not sufficient to predict the domain in order to correctly verify the target sense.

If such in-domain ambiguities were not applicable for a given target sense, the discovery of examples that applied ***mixing of contexts*** formed another strategy of generating hard examples. In these instances, an out-of-domain sense is used in an in-domain context or vice-verse, for example, in the context “If we now look at an object, for example an apple, the light rays reflected from this *apple* hit our cornea”, the target *apple* –the fruit– is used in the context of the eyeball, which in German translates to “eye *apple*”.

Finally, the notion of ***trigger words*** was introduced, where not the entire context, but only specific words draw the connection to the target sense from a different domain. An example would be “Rabbits also like the *flower* with the intense scent.”, where the term *flower* –describing the plant– could also be used to refer to the tail of rabbits.

## 4.3. Human Baseline

For retrieving an upper bound of performance for the presented dataset, we created a *Human Baseline* from a sub-sample of the test set. Herefore, 250 instances were randomly selected and split into two batches with 150 instances each, resulting in an overlap of 20%. Each of these batches was assigned to an annotator. The annotators –being German native-speakers and non-experts in all of the four focus domains– were instructed to solve the dataset without the usage of external knowledge sources (e.g., when they are not familiar with the meaning of the target sense).

A summary of the human performance evaluation can be found in Table 4. The overall performance with an average of 90.3% accuracy is remarkably high, with individual scores of 90% and 90.7%. Also the inter-rater agreement –calculated on the overlapping examples– denoted by a Cohen’s Kappa statistic of 0.84 shows that the task is relatively straight-forward and solvable for humans.

When analysing the different subsets, i.e., domain-independent and domain-bound, we can see that domain-independent instances seem harder to solve

[11] https://www.duden.de

[12] Unfortunately, we were not able to derive the information on sense frequency from any other source.

| **subset** | ***Target Sense Commonality*** | | ***Hard Example Strategy*** | | | |
|---|---|---|---|---|---|---|
| | **1st$_{\text{WIK}}$** | **any$_{\text{WIK}}$ / any$_{\text{Duden}}$** | **ambig.** | **mixing** | **trigger** | **none** |
| FOOD | 24.1% | 81.4% / 85.5% | 9.7% | 4.8% | 2.8% | 82.8% |
| HUNT | 4.3% | 70.0% / 57.1% | 13.6% | 9.3% | 10.7% | 66.6% |
| MED | 50.7% | 90.0% / 100% | 2.1% | 2.1% | 6.4% | 89.3% |
| ZOO | 71.5% | 92.9% / 96.1% | 11.0% | 2.6% | 7.1% | 79.2% |

Table 3: Qualitative characteristics of domain-bound instances in terms of strategies applied to create hard examples and commonality of the target sense; *ambig.* stands for in-domain ambiguity, *mixing* for mixing of contexts, *trigger* for trigger word. For the target sense commonality, presented are the percentage of instances whose target sense is listed as (*1st$_{WIK}$*): the first sense in Wiktionary and *(any$_{WIK}$/any$_{Duden}$)*: any sense in Wiktionary or Duden

than domain-focused ones, with an average accuracy of 83.6 (with individual scores of 81.6 and 85.5) compared to 97.3 (98.4 and 95.9). This phenomenon is in line with previous studies and can be explained by the fact that general-domain meanings are often closer together, only having nuanced differences. Therefore, it is harder to evaluate whether the given target sense fits, or if there could be another meaning that is even closer. This is also reflected in the high recall for domain-independent examples.

Domain-specific meanings on the other hand are often more clearly distinguishable, resulting in higher annotation scores. Still, the evaluation performance of human annotators on the domain-bound examples is surprisingly high, especially on domains with highly specialised language such as *Hunting*, where all examples were correctly annotated. The lowest domain-focused annotation score was achieved in the domain of *Medicine* with an average accuracy of 94.7, resulting from the individual scores of 100 and 89.5.

## 5. Experimental Setup

To establish a baseline for the introduced dataset, we evaluate the performance of a basis model in two different settings: *monolingual* and *cross-lingual*.

For our experiment we considered HyperBERT3, a classifier based on a pre-trained contextualised BERT model, as our baseline model to fine-tune on our dataset. First, we apply a weak supervision mechanism described in (Huang et al., 2019) to our input data, by surrounding the target word in the context with a special symbol (in this work, we used *$*). As shown in the original paper, this weak supervision mechanism helps the model identifying the target token. Then, the definition and hypernyms are concatenated –separated by the same special symbol– and fed together with the context into the BERT model. The resulting representations (i.e., for the [CLS] token, target tokens, and the concatenated sense descriptors) are fed into a top-level classifier to create the final predictions. The implementation of HyperBERT3 can be found at `https://github.com/semantic-web-company/wic-tsv/blob/master/HyperBert/HyperBert3.py`.

For the monolingual setting, we used HyperBERT with *bert-base-german-cased* as the base model, which we fine-tuned and evaluated on WiC-TSV-de. For the cross-lingual setting, a model based on *bert-base-multilingual-cased* was fine-tuned on the English WiC-TSV training and development set, and evaluated on WiC-TSV-de. For each of the two settings, we fine-tuned the HyperBERT3 model for 8 epochs, and chose the best performing model. For the final results, we calculated the average of three runs.

## 6. Results

**Monolingual Setting** An overview of the performance of HyperBERT3 in monolingual setting can be seen in Table 4. The overall performance remains with an $F_1$ score of 71% relatively low, especially in terms of precision. This underlines the challenging nature of the dataset, as the baseline model leaves a big gap of almost 20 percent points to human annotation performance. For comparison, the gap between the best-performing model and human performance on the original English WiC-TSV dataset was 8.6 percent points.

The domain-independent instances –despite their relative coarse-granularity– pose a major challenge to the TSV model, which was only able to achieve an average of 71.7% $F_1$ on this subset. However, the performance achieved on these instances is with 13.8 percent points difference closest to the human annotation performance, compared to the other subsets.

While the average performance of domain-independent and domain-bound instances are similar, there are some differences between the different domains. The most difficult domain seems to be *Hunting* with an average $F_1$ score of 62%, followed by *Zoology*. On the most successful subset, *Gastronomy*, HyperBERT3 reached an $F_1$ score of 76%. These performance difference could be partly explained by the qualitative characteristics of the domain-bound subsets: the *Hunting* and *Zoology* domain contain a greater number of in-domain ambiguous target words and context-mixing instances, leading to examples where the correct label could not be guessed from identifying the domain of the whole context. Interestingly, these types of instance seem to cause less erroneous predictions in the *Gastronomy* domain, which contains more of these hard examples than the medical domain, but shows a better performance.

Another interesting aspect that arises from this analysis is that the performance on all subsets is leaning towards

| | **WiC-TSV-de** | | | |
|---|---|---|---|---|
| | **Acc** | **Prec** | **Rec** | **F1** |
| HBERT$_{\text{de}}$ | 68.5 ± 0.3 | 65.8 ± 0.6 | 77.7 ± 2.8 | 71.2 ± 0.9 |
| HmBERT$_{\text{en\_de}}$ | 56.8 ± 0.4 | 53.9 ± 0.4 | 94.7± 2.4 | 68.7 ± 0.4 |
| All-True | 50.2 | 50.2 | 100 | 66.82 |
| *Human* | *90.3* | *87.7* | *94.9* | *91.1* |
| | *WIK* | | | |
| | **Acc** | **Prec** | **Rec** | **F1** |
| HBERT$_{\text{de}}$ | 67.9 ± 0.6 | 64.3 ± 0.1 | 81.1 ± 3.5 | 71.7 ± 1.3 |
| HmBERT$_{\text{en\_de}}$ | 56.3 ± 0.7 | 53.7 ± 0.5 | 94.8 ± 2.2 | 68.5 ± 0.5 |
| All-True | 50.2 | 50.2 | 100 | 66.8 |
| *Human* | *83.6* | *78.7* | *93.7* | *85.5* |
| | *FOOD* | | | |
| | **Acc** | **Prec** | **Rec** | **F1** |
| HBERT$_{\text{de}}$ | 74.9 ± 1.4 | 74.0 ± 3.8 | 79.0± 11.6 | 75.7 ± 4.0 |
| HmBERT$_{\text{en\_de}}$ | 59.1 ± 2.5 | 55.7 ± 1.9 | 92.7 ± 3.9 | 69.5 ± 1.1 |
| All-True | 50.3 | 50.3 | 100 | 67.0 |
| *Human* | *97.3* | *100* | *95.8* | *97.8* |
| | *HUNT* | | | |
| | **Acc** | **Prec** | **Rec** | **F1** |
| HBERT$_{\text{de}}$ | 65.2 ± 4.5 | 67.2 ± 2.4 | 58.6± 12.3 | 62.1 ± 8.1 |
| HmBERT$_{\text{en\_de}}$ | 57.1 ± 0.6 | 54.2 ± 0.3 | 93.3 ± 4.7 | 68.5 ± 1.3 |
| All-True | 50.0 | 50.0 | 100 | 66.7 |
| *Human* | *100* | *100* | *100* | *100* |
| | *MED* | | | |
| | **Acc** | **Prec** | **Rec** | **F1** |
| HBERT$_{\text{de}}$ | 72.2 ± 2.1 | 68.8 ± 2.6 | 81.4 ± 1.2 | 74.5 ± 1.2 |
| HmBERT$_{\text{en\_de}}$ | 61.0 ± 1.7 | 56.3 ± 1.1 | 98.6 ± 2.0 | 71.6 ± 0.7 |
| All-True | 50.0 | 50.0 | 100 | 66.7 |
| *Human* | *94.7* | *100* | *92.9* | *96.2* |
| | *ZOO* | | | |
| | **Acc** | **Prec** | **Rec** | **F1** |
| HBERT$_{\text{de}}$ | 64.3 ± 0.3 | 62.0 ± 2.4 | 77.8± 13.3 | 68.2 ± 4.1 |
| HmBERT$_{\text{en\_de}}$ | 52.0 ± 1.6 | 51.3 ± 0.9 | 93.6 ± 1.0 | 66.3 ± 0.5 |
| All-True | 50.3 | 50.3 | 100 | 67.0 |
| *Human* | *97.4* | *95.0* | *100* | *97.4* |

Table 4: Test set performance of the monolingual (HBERT$_{\text{de}}$) and cross-lingual (HmBERT$_{\text{en\_de}}$) HyperBERT3 baseline models, an All-True classifier, and human annotators on WiC-TSV-de, in terms of accuracy, precision, recall, and $F_1$ for the positive label. For the models, the mean performance and standard deviation out of three runs is presented. The overall performance (top) is further split into the performances on the domain-independent (WIK) and the domain-bound subsets (FOOD, MED, HUNT, and ZOO).

recall, except for the *Hunting* domain. A possible explanation for this could be the aforementioned highly specialised, yet ambiguous language used in this area, which results in low commonality of the target senses. However, it should also be noted that the standard error of the recall is quite high for most domains.

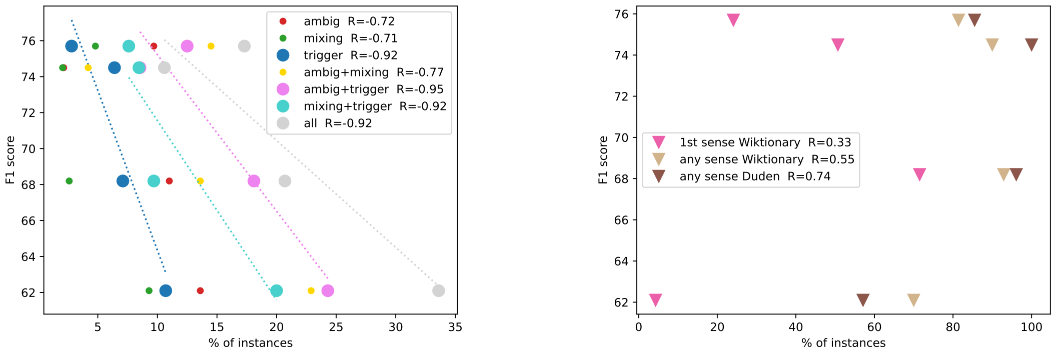
**Cross-Lingual Transfer Setting** The outcomes of the cross-lingual transfer setting are also to be found in Table 4. It can be seen that the overall performance is very low, being only 1.9 percent points above the $F_1$ score of the naive All-True lower-bound baseline. The resulting classifier indeed is close to an all-true classifier, as not even 13% of its predictions on the entire test set were *False* labels. Generally, the incorporated multilingual BERT model seems to have difficulties to adapt to the presented task, as the model also performed quite poorly in a monolingual setting, i.e., when both fine-tuned and tested on the WiC-TSV-de, reaching an accuracy of only 61% (Precision: 62.7, Recall: 54.6). When comparing the different domain-bound subsets, a similar image than in the monolingual analysis is drawn: while the best performance was achieved on *Medicine*, and *Gastronomy*, with 61.0% and 59,1% accuracy, the *Zoology* subset is the most difficult one. Interestingly, the performance difference of the multilingual model on *Zoology* and *Hunting* is notable bigger than is is in the monolingual equivalent.

**Domain-difficulty** To estimate the effect of challenging examples on the performance of the TSV models we calculated the Pearson correlation of $F_1$ scores and the extent to which a certain strategy (or combinations of) contributed to the creation of the domain-bound subsets (cf Table 3). A summary of the is presented in Figure 1.

For the monolingual HyperBERT3 model, we identify that, overall, all *Hard Example Strategies* show an expected negative correlation (Pearson coefficient between -0.71 and -0.95) and that the correlation increases when combining different strategies. The strongest correlation was observed on the sum of *in-domain ambiguous* and *trigger word* examples.

Curiously, the analysis on the *Target Sense Commonality* suggests a weaker connection between the performance of the tested models and the usage of uncommon target senses. While the correlation between performance and percentage of instances whose target sense is listed as the first sense in Wiktionary is only 0.33, the Pearson coefficient when comparing whether the target sense is listed at all ranges from 0.55 to 0.74, depending on the resource compared against. Interestingly, the correlation to the expert-curated Duden is much higher than to the crowd-sourced Wiktionary, which could indicate that expert-curated sources are a more reliable source regarding sense commonality.

Due to the general low performance of the model, the analysis of multilingual HyperBERT3 in the cross-lingual setting, yield to inconclusive outcomes. While the correlation coefficient for all types of *Hard Example Strategies* and their combinations are negative, the correlations are weaker than for the monolingual setting, with Pearson coefficients between -0.16 (mixing) and -0.77 (ambig). For the *Target Sense Commonality* analysis, the coefficients range from -0.24 (first sense in Wiktionary) to 0.18 (any sense in Duden).



(a) Hard Example Strategies

(b) Target Sense Commonality

Figure 1: Scatter plots for $F_1$ scores on the different domain-bound subsets in the monolingual setting and (a) percentage of hard examples originated from different strategies, (b) percentage of target senses with different commonality. For those combinations, where the absolute Pearson coefficient is greater than 0.8, a trend-line is added.

Nevertheless, from this analysis it is observable that the introduced strategies for collecting contexts indeed lead to hard instances. It shall be noted, however, that the number of datapoints of these correlation analysis is quite low and therefore the presented correlation values rather represent an indicated trend than reliable statistics.

## 7. Conclusions

In this paper, we introduced WiC-TSV-de, a German dataset for evaluating the disambiguation capabilities of Target Sense Verification models. Due to its binary formulation, the dataset allows for the development of flexible models that are independent of sense inventories. While the training and development set consist of instances that are not bound to any specific domain only, the test set also contains domain-bound instances originating from fours different domains, being *Gastronomy*, *Medicine*, *Hunting*, and *Zoology*, allowing to evaluate the domain transfer capabilities of models. These domain-bound instances incorporate adversarial examples that exploit in-domain ambiguities, mixing of contexts and trigger words, to make the dataset more challenging. Nevertheless, human annotation performance of over 90% underlines the overall clarity of the test instances. By combining WiC-TSV-de with the original English WiC-TSV, the dataset can further be used to analyse the cross-lingual transfer capabilities of disambiguation models.

Initial experiments on WiC-TSV-de showed that the BERT-based baseline model was not able to perform satisfyingly on the dataset, leaving almost 22 percent points gap to human performance. In general, domains with more adversarial examples appear to be harder for the model. When evaluating in cross-lingual setting, the baseline model showed even worse performance of only approximately 57% accuracy, as the model adapted an all-true-classifier-like behaviour.

This dataset therefore opens up avenues for the development of models that can successfully perform domain- and language-transfer, and thus are capable of performing well in a variety of different settings without the need of large amounts of scenario-specific training data.

## 8. Bibliographical References

Agirre, E., Magnini, B., de Lacalle, O. L., Otegi, A., Rigau, G., and Vossen, P. (2007). Semeval-2007 task 01: Evaluating wsd on cross-language information retrieval. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, page 1–6, USA. Association for Computational Linguistics.

Agirre, E., De Lacalle, O. L., Fellbaum, C., Marchetti, A., Toral, A., and Vossen, P. (2010). Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 123–128.

Duygu Ataman, et al., editors. (2021). *Proceedings of the 1st Workshop on Multilingual Representation Learning*, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Breit, A., Revenko, A., Rezaee, K., Pilehvar, M. T., and Camacho-Collados, J. (2021). WiC-TSV: An evaluation benchmark for target sense verification of words in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645, Online, April. Association for Computational Linguistics.

Faralli, S. and Navigli, R. (2012). A New Minimally-supervised Framework for Domain Word Sense Disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language*

*Processing and Computational Natural Language Learning*, pages 1411–1422, Jeju, Korea.

Gella, S., Elliott, D., and Keller, F. (2019). Cross-lingual visual verb sense disambiguation. 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 1998–2004, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Henrich, V. and Hinrichs, E. (2012). A comparative evaluation of word sense disambiguation algorithms for German. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 576–583, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Huang, L., Sun, C., Qiu, X., and Huang, X. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3507–3512, Hong Kong, China, November. Association for Computational Linguistics.

Okumura, M., Shirai, K., Komiya, K., and Yokono, H. (2010). Semeval-2010 task: Japanese wsd. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, page 69–74, USA. Association for Computational Linguistics.

Pasini, T. (2020). The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4936–4942. International Joint Conferences on Artificial Intelligence Organization, 7. Survey track.

Procopio, L., Barba, E., Martelli, F., and Navigli, R. (2021). Multimirror: Neural cross-lingual word alignment for multilingual word sense disambiguation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3915–3921. International Joint Conferences on Artificial Intelligence Organization, 8. Main Track.

Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of EACL*, pages 99–110, Valencia, Spain.

Sauermost, R. and Freudig, D. (1998). Lexikon der Biologie : in vierzehn Bänden.

Scarlini, B., Pasini, T., and Navigli, R. (2020). Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5905–5911, Marseille, France, May. European Language Resources Association.

Vial, L., Lecouteux, B., and Schwab, D. (2018). UFSAC: Unification of Sense Annotated Corpora and Tools. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

## 9. Language Resource References

Breit A. et. al. (2021). *WiC-TSV (Word-in-Context Target-Sense-Verification) Dataset*. Association for Computational Linguistics.

Christiane Fellbaum. (1998). *WordNet: An Electronic Database*. MIT Press.

Hamp, Birgit and Feldweg, Helmut. (1997). *GermaNet - a Lexical-Semantic Net for German*.

Liu, Qianchu and Ponti, Edoardo Maria and McCarthy, Diana and Vulić, Ivan and Korhonen, Anna. (2021). *AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples*. Association for Computational Linguistics.

Martelli, F., Kalach, N., Tola, G., and Navigli, R. (2021). SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online, August. Association for Computational Linguistics.

Pilehvar, Mohammad Taher and Camacho-Collados, Jose. (2019). *WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations*. Association for Computational Linguistics.

Raganato, Alessandro and Pasini, Tommaso and Camacho-Collados, Jose and Pilehvar, Mohammad Taher. (2020). *XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization*. Association for Computational Linguistics.

Siegel, Melanie and Bond, Francis. (2021). *OdeNet: Compiling a GermanWordNet from other Resources*. Global Wordnet Association.