Bachelor’s Thesis

Creating and Evaluating a New Specificity Metric based on WordNet

Author
Darja Jepifanova
darja.jepifanova@student.uni-tuebingen.de

Supervisor
Çağrı Çöltekin
cagri.coeltekin@uni-tuebingen.de

A thesis submitted in partial fulfilment
of the requirements for the degree of

Bachelor of Arts
in
International Studies in Computational Linguistics

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

October 2024

Antiplagiatserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst habe, dass ich keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe, dass ich alle wörtlich oder sinngemäß aus anderen Werken übernomme- nen Aussagen als solche gekennzeichnet habe, dass die Arbeit weder voll- ständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsver- fahrens gewesen ist, dass ich die Arbeit weder vollständig noch in wesentlichen Teilen bereits veröffentlicht habe, dass das in Dateiform eingereichte Exemplar mit dem eingereichten gebundenen Exemplar übereinstimmt.

Depifanova

Tübingen, den 1. Oktober 2024 Darja Jepifanova

Abstract

In computational linguistics, specificity quantifies how much detail is engaged in text. It is useful in many NLP applications such as summarization and information extraction. It can also be used as a text quality metric, which can provide a more transparent evaluation of machine-generated text. Yet to date, expert-annotated data for sentence-level specificity are scarce and confined to the news or social media genre. In addition, systems that predict sentence specificity are classifiers trained to produce binary labels (general or specific).

In this thesis, I introduce a new specificity metric based on WordNet, which posits that lower synsets in the semantic hierarchy represent more specific concepts. Based on this principle, I trained a Siamese network to distinguish between specific and general sentences based on the depth of the corresponding synsets. The evaluation of the resulting continuous specificity scores involved statistical analysis, comparisons with existing metrics, and human evaluations. The analysis revealed a Pearson correlation of 0.38 with a Twitter-based dataset containing annotated specificity, suggesting promising outcomes. However, human evaluations resulted in a correlation of just 0.19 and low inter-annotator agreement, highlighting the metric's limitations. Despite these challenges, this study provides a foundation for future enhancements and applications in natural language generation, aiming to improve the quality and precision of machine-generated texts.

Contents

List of Tables

List of Figures

1 Introduction

Large Language Models (LLMs) have recently demonstrated extraordinary capabilities in a range of natural language processing (NLP) tasks, including language translation, text generation, and question answering. Their advancements have dramatically transformed the field of NLP, making significant improvements in model performance and applicability. Though this success of LLMs has prompted a substantial increase in research contributions, such rapid growth has made it difficult to have a clear overview of the field and understand the overall impact of these improvements (Raiaan et al., 2024).

Evaluating language models is complex due to the absence of a universally accepted standard. Unlike tasks with clear-cut answers, language processing often involves multiple valid outputs, making it hard to define a "correct" response. This subjectivity results in diverse evaluation criteria and a lack of consensus on metrics, which continues to hinder progress.

Common metrics for evaluating generated text, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), focus on surface-level structure by assessing n-gram overlap with a reference text. BERTScore (Zhang et al., 2020), on the other hand, measures semantic similarity by comparing contextual embeddings, providing a more meaning-oriented evaluation. However, all these metrics rely on a reference text as the gold standard, which can be problematic since different human-written texts may be equally correct or valuable in various ways.

While these metrics are useful for measuring surface-level similarity, they fall short in assessing deeper qualities such as content accuracy, coherence, informativeness, and readability (Nguyen et al., 2024; Fabbri et al., 2021). Human evaluations, while insightful, are resource-intensive and variable. Additionally, as models and applications evolve, so do the standards for evaluation, requiring ongoing development of new metrics and methodologies. These limitations highlight the need for more sophisticated evaluation methods that can capture the true quality of generated summaries.

Recent explorations of LLM-based NLG evaluation reveal new challenges. Hu et al. (2024) observed that evaluation results from LLMs for one aspect can achieve a higher correlation with human judgments on a different, unrelated aspect. Additionally, correlations between LLM-generated scores across different aspects are significantly higher than those between human judgments. These findings raise concerns about the reliability of LLM evaluations, as they suggest that LLMs may confuse different aspects of text quality, leading to misleading evaluation results.

Quality of the generated text is more than similarity to some gold standard text for the specific prompt, which sometimes is not even reviewed by humans. A better approach for evaluations would be more complex and detailed analysis of the individual aspects that affect the overall quality of the text. Recently, this has become the dominant paradigm in human evaluation (Fabbri et al., 2021). Text quality does not involve a single aspect but is a combination of numerous and diverse criteria including spelling, grammar, organization, informative nature, creative and beautiful language use and specificity (Louis, 2013).

These aspects, however, differ based on the interest of concrete research and text generation task itself. For example, text summarization typically uses four dimensions for evaluation: coherence, consistency, fluency, and relevance (Fabbri et al., 2021). One of the most popular criteria which is reviewed for various NLG tasks is informativity. Of course, it also relevant to consider in tasks like summarization, where the goal is to condense a large body of information into a concise format without losing essential details. The definition of informativity varies in the literature, but we could define it as the amount of useful information in the text (Novikova et al., 2017). It refers to the richness of content and factual density of the text (Lex et al., 2012). We should also note that this measure is highly related, but not to be confused, with the length of the response. Logically, more text convey more information Levshina (2022).

However, informativeness alone does not guarantee that the content is well-targeted or sufficiently detailed for the task at hand. This is where specificity comes into play. Specificity refers to the amount of precision and detail provided in the text. While informativity ensures that the content is valuable, specificity ensures that this value is delivered with the necessary depth and precision. To illustrate a difference between these two metrics, here is an example of two sentences. Both of them are informative, but sentence b) is more specific:

(a) The weather has been quite pleasant recently.

(b) Over the past week, the average temperature has been a comfortable 22 degrees Celsius with sunny skies and low humidity.

Both sentences express that the weather is good, but the first is vague, offering only a general sense of pleasantness. The second sentence, however, provides specific details about temperature, sky conditions, and humidity levels, making the description more precise and informative.

Specificity is a crucial aspect of various NLG tasks, such as automatic summarization, question answering, and chatbot responses. Despite advancements in LLMs like GPT-3 and GPT-4, these models frequently generate filler sentences that lack informativeness, often because they prioritize fluency over content, leading to responses that are grammatically correct but vague.

Beyond NLG tasks, specificity is also crucial in information retrieval, especially when integrated with Named Entity Recognition (NER). By identifying and retrieving more precise and relevant documents, this combination can enhance modern techniques like Retrieval-Augmented Generation (RAG) and active learning, which rely on accessing highly targeted information to improve model outputs.

Morevoer, in low-resource settings, where annotated data is scarce, selecting highly specific sentences for annotation can significantly improve the training of NLP models, maximizing the informativeness of each labeled instance. Specificity also finds applications beyond computational contexts, such as in political discourse, argumentation, classroom discussions, clinical psychology and sociolinguistics (Gao et al., 2019).

The previous research relied on sentence specificity prediction systems that are trained to produce 2-3 categories of specificity sentences (Li and Nenkova, 2015; Louis and Nenkova, 2012). Furthermore, the few available open-source tools are constrained by the nature of their training data: models like the one developed by Li and Nenkova (2015) were trained on news sentences, while Gao et al. (2019) based their work on a Twitter dataset. As a result, these models often fail to generalize well across different text genres, limiting their broader applicability in NLP tasks. For a more comprehensive discussion of related work, please refer to Chapter 2.

To overcome the limitations of current specificity prediction methods, this thesis introduces a new specificty metric based on WordNet, a comprehensive semantic network for English developed by George Miller (Miller and Charles, 1991). In this lexical database words are organized hierarchically through hypernym (“IS A”) relationships. Higher synsets represent more abstract concepts, while lower synsets denote more specific ones.

Based on this structure, I constructed a training dataset of sentence pairs derived from synsets at different depth in the WordNet hierarchy. Using this data, I trained a Siamese neural network with BERT embeddings to predict a specificity score between 0 and 1 for each sentence, indicating its level of specificity. The model’s output is a binary comparison that determines which of two sentences is more specific. But the primary focus is on its ability to generate a continuous specificity score for any given sentence. More details on the data and model architecture are discussed in Chapters 3 and 4.

Evaluating the developed specificity metric was the most challenging aspect of this work. The evaluation process involved conducting descriptive statistics of the specificity scores, comparing my metric with several publicly available specificity scorers through correlation analysis and assessing its alignment with human judgments. Detailed results and methodologies of the evaluation process are discussed in Chapters 3.3 and 5.

2 Related Work

2.1 Defining Specificity

Sentences vary in the level of details they convey. Specificity can be defined as the extent to which sentences convey detailed and concrete content (Li and Nenkova, 2015). To better understand specificity, it is essential to identify its key characteristics, as discussed in the literature. This section will summarize the various dimensions of specificity, including its relationship with word frequency, syntactic features, named entities mentions, and other relevant factors.

Specificity is closely related to the frequency of words, with inverse document frequency (IDF) being a key measure (Ko et al., 2018; Li and Nenkova, 2015; Louis and Nenkova, 2012; Yao et al., 2016). Specific sentences are more likely to contain rare or less frequent words, indicating detailed and unique content. Respectively, words with high IDF values, which occur less frequently, are considered more specific (Yao et al., 2016). Similarly, Zhang et al. (2018) evaluates specificity in generated responses by calculating distinct unigrams and bigrams, suggesting that more unique word combinations indicate higher specificity and diversity. Building on this, bigram and trigram language models extend the idea by evaluating the likelihood of word sequences in context rather than just individual word frequency. Higher perplexity in these models indicates more specific and detailed content (Louis and Nenkova, 2012).

One fundamental indicator of specificity is sentence length. Generally, specific sentences tend to be longer, this is reflected in features like the number of words in a sentence and the average length of words (Gao et al., 2019; Ko et al., 2018; Li and Nenkova, 2015; Louis and Nenkova, 2012). Syntactic features such as the count of verbs, nouns, adjectives, adverbs, and complex phrases also play a role in identifying specificity. Specific sentences often feature longer verb phrases, more nouns and complex syntactic structures, while general sentences tend to include more adjectives and adverbs, and have simpler phrase structures (Louis and Nenkova, 2012; Gao et al., 2019).

Numerical mentions and named entities further indicate specificity (Gao et al., 2019; Ko et al., 2018; Li and Nenkova, 2015; Louis and Nenkova, 2012). It is described as the number of specific entity names, numerical values, and times/dates scaled by the overall word count of a document Jiang and Srinivasan (2023).

Sentences with more punctuation (e.g., commas, colons, parentheses) and explicit discourse connectives (words that connect clauses or sentences logically) tend to be more specific. This reflects how specific sentences often contain multiple clauses or more elaborated information (Ko et al., 2018; Li and Nenkova, 2015).

Another important factor is polarity. General sentences frequently express strong opinions or sentiments, which can be quantified by the presence of positive, negative, and polar (non-neutral) words. Discourse connectives like “but” and “however” are also common in general sentences, while specific sentences tend to be less subjective and rely more on direct information (Ko et al., 2018; Louis and Nenkova, 2012). To further explore the role of emotion in specificity, Gao et al. (2019) examine emotional features expressed not only through positive or negative words but also through the use of emojis.

The distribution of specific and general sentences within a text follows a recognizable pattern. For instance, technical writing often exhibits an hour-glass structure: the introduction begins with general content, narrows down to specific details in the methods and results sections, and then broadens again in the conclusion. A similar pattern can be observed on a smaller scale in summaries, where the opening often provides a general overview, while the body contains specific details and the conclusion may generalize the information again (Louis and Nenkova, 2012).

2.2 Methods for Measuring Specificity

In recent years, several methods have been developed to predict sentence specificity. Li and Nenkova (2015) proposed a semi-supervised model that utilized both labeled and unlabeled data to train their model. The labeled data was sourced from the Penn Discourse Treebank (PDTB). In particular, the researchers concentrated on the Instantiation and Specification relations, where a general sentence is often succeeded by a more specific one that offers additional details or examples. For instance, in a pair of sentences like “He spent a lot of money on his art business” (general), followed by “Last week, he spent $23 million at an auction on seven artworks, including a Picasso”(specific), the first sentence serves as a broad claim, while the second one offers specific details that instantiate the idea.

The model employed a range of features to predict sentence specificity, focusing on surface properties like sentence length and syntactic complexity, as well as lexical features that captured more fine-grained distinctions. The authors also tackled the issue of lexical sparsity in small datasets by using more general, non-sparse word representations such as Brown clusters and neural word embeddings. They employed a simple logistic regression classifier to predict sentence specificity based on these features. Their model, trained on this rich dataset of general and specific sentences, was evaluated against human-labeled test data and showed competitive performance with prior works.

Building on this approach, Li and Nenkova also adopt a co-training method in a semi-supervised learning framework. In co-training, two independent classifiers are trained on a labeled dataset, leveraging different views of the data to make decisions about class labels. This process is iterative: each classifier labels a large number of unlabeled examples, and the most confidently labeled instances are added to the training set for the other classifier. This methodology not only addresses the issue of limited labeled data but also enhances model robustness by incorporating the insights from both classifiers.

Louis and Nenkova (2012) developed a classifier to distinguish between general and specific sentences in news articles also based on existing discourse annotations in PDTB. They utilized same two types of discourse relations: Instantiation and Specification. The classifier was trained using various features that captured lexical and syntactic information, including counts of adjectives, adverbs, and verb phrases, as well as word specificity and polarity. Additionally, language models were employed to assess the likelihood of word transitions, helping to differentiate between expected and unexpected content in sentences.

The study by Gao et al. (2019) collect and annotate on specificity a dataset of over 7,000 tweets. For the annotation process, the authors adopted a fine-grained scale from 1 to 5, with detailed guidelines provided to ensure consistency among annotators. The annotation was conducted using Amazon Mechanical Turk.

With the obtained data they trained a regression model to predict the specificity of tweets, treating specificity as a continuous value rather than categorizing it into discrete classes. They specifically employed Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel. To the best of our knowledge, this study presents the largest open-source dataset with annotated specificity and the only specificity classifier that generates continuous scores instead of binary labels.

2.3 WordNet: Applications and Limitationst

WordNet is a large semantic database for English that organizes words into sets of synonyms called “synsets”, which capture distinct meanings or senses of a word. Each synset represents a unique concept and is interconnected with other synsets through semantic relations like hypernymy and hyponymy(Miller and Charles, 1991). One of its key advantages is its focus on the semantic relationships between words rather than syntactic structures. This allows WordNet to be especially useful in word-sense disambiguation, where the goal is to assign the most contextually appropriate sense to a word in a given text (Navigli, 2009). WordNet’s ability to represent fine-grained distinctions between meanings and its hierarchical structure make it ideal for computational tasks that require an understanding of semantic depth and similarity.

WordNet has also been widely employed to measure semantic similarity and semantic density. For example, Resnik (1995) created a measure of semantic similarity in an “IS A” taxonomy that leverages information content to assess how much detail a word carries. In this approach, words that are lower in the taxonomy are considered more informative. Meng et al. (2014) extend this idea by introducing a metric that incorporates both path information (how far apart two concepts are in the taxonomy) and information content (how informative a word is based on its frequency and position). A similar principle of depth can be applied to specificity, as concepts lower in the hierarchy tend to be more detailed and specific.

The only previous application of WordNet depth in specificity prediction comes from the work of Louis and Nenkova (2012), who utilized it as one of the features in their specificity classifier. Building on the idea that more specific sentences are likely to contain more specific words and details, they leveraged hypernym relations from WordNet to compute the specificity of individual words. For each noun and verb in a sentence, they recorded the length of the path from the word to the root of the WordNet hierarchy through its hypernym relations. They computed the average, minimum, and maximum distances for nouns and verbs separately and used these as features in their model.

Despite its widespread use, WordNet has several notable limitations. One of the most widely discussed issues is that certain semantic relations are more suited to concrete concepts than to abstract ones. For instance, it is relatively straightforward to establish hyponym/hypernym relationships for tangible entities, such as “dog” being a type of “animal”. However, classifying abstract concepts like “justice” or “freedom” into well-defined hierarchical relationships proves challenging (Rudnicka et al., 2018).

Furthermore, the organization of concepts within WordNet can often be perplexing. The irregular densities of links between concepts lead to unexpected measures of conceptual distance. For example, two concepts that might intuitively belong at the same level of abstraction can be found at different depths in the hierarchy. A case in point is “horse”, which is situated 10 levels from the root (considered as “entity”), while “cow” is positioned 13 levels deep (Richardson et al., 1994). This inconsistency can complicate the measurement of specificity, as the hierarchical arrangement does not always reflect the inherent relationships between concepts.

Additionally, the imbalance in the WordNet hierarchy is significant. While some concepts have deep, well-defined hierarchies, many others do not, leading to uneven distributions of specificity across different words. This lack of balance can hinder the effectiveness of specificity measurements, as words at similar levels of abstraction might not provide equally detailed information.

3 Data

3.1 Raw Sentences

I selected the WordNet lexical database for training data because specificity is more closely tied to the semantic richness of word senses rather than their syntactic structure. By focusing on word senses rather than individual words, I aimed to better handle word ambiguity. For instance, the adjective “light” in the phrase “a light heart” can mean either “irresponsible” or “not heavy”. In this case, the latter interpretation is more concrete compared to the idiomatic usage. Another issue is with words that can function as different parts of speech, as the same word form can have multiple, distinct meanings depending on its syntactic role. For instance, “break” as a verb might mean to fracture an object, while “break” as a noun might refer to a pause or interval. Intuitively, the verb usage of “break” conveys a more specific meaning compared to the more abstract noun usage.

As training data, I used example sentences of synsets since they can provide the necessary context for understanding the specificity of a word sense. This approach also aligns with the standard practice in the field, where evaluating specificity at the sentence level is more useful for downstream applications. Real-world NLP tasks – such as summarization, question answering, and information extraction – rely on entire sentences rather than isolated words.

The WordNet database, accessed via NLTK (Loper and Bird, 2002), includes approximately 150,000 synsets, of which only about 25,000 have example sentences. Specifically, there are 8,742 noun synsets, 9,691 verb synsets, 4,355 adjective synsets, and 3,192 adverb synsets. Adjectives and adverbs were excluded from the analysis due to their lack of hypernym relations. For the dataset, I sampled 8,000 pairs of nouns and 9,000 pairs of verbs from the same hierarchical path, keeping track of depth and identifying which synset was more specific by binary labels – 0 for the first sentence being more specific and 1 for the second. To ensure balanced representation, I evenly distributed the labels and shuffled the dataset.

Let’s consider a typical example to better understand the structure of the data. In Table 1, the row with ID `12070` provides a pair of synsets: Synset(’communicate.v.01’) and Synset(’buzz.v.04’). The corresponding example sentences are “Please communicate this message to all employees” and “He buzzed the servant”. The synset “communicate” appears at depth 4 in the WordNet hierarchy, while “buzz” is at depth 11. The label of 1 indicates that “He buzzed the servant” is considered more specific than “Please communicate this message to all employees”. This distinction is evident as “buzzed” describes a more precise and context-specific action compared to the more general act of “communicating”. Notably, the more specific sentence is not necessarily longer.

id	synset1	depth1	sentence1	synset2	depth2	sentence2	specific
12070	’communicate.v.01’	4	Please communicate this message to all employees	’buzz.v.04’	11	He buzzed the servant	1
14773	’gore.v.02’	5	gore a skirt	’cut.v.07’	4	cut a dress	0
238	’clemency.n.02’	14	He threw himself on the mercy of the court	’quarter.n.14’	15	He surrendered but asked for quarter	1
2146	’bad_luck.n.02’	6	If I didn’t have bad luck I wouldn’t have any luck at all	’luck.n.02’	5	Bad luck caused his downfall	0
4	’street.n.05’	8	It is a friendly street	’neighborhood.n.02’	7	It is a friendly neighborhood	0
16723	’act.v.04’	3	She acts as the chair	’criticize.v.02’	4	She criticize as the chair	1

Table 1: Examples of general and specific WordNet sentence pairs

After quick inspection of the dataset, we can see that some example sentences consist of noun or verb phrases rather than full grammatical sentences. For instance, entry with ID `14773` contains two verb phrases which were not filtered out. This variability in the structure of the training data can affect the model’s performance when applied to full sentences.

It is also worth noting that the majority of depth differences between sampled synsets are either -1 or 1, suggesting that the actual difference in specificity levels between corresponding sentences is often not substantial. This phenomenon is primarily due to the structure of WordNet, which is organized into a relatively shallow hierarchy with numerous direct children and few deeply nested descendants. As a result, there is a high frequency of synsets that are close in depth, leading to most sampled pairs having minimal depth differences. However, we cannot confidently assert that sentence pairs with the same depth difference are comparably distinct in their concreteness for the reasons was discussed in Chapter 2.3.

3.2 Sentences with the Same Context

To enhance control over the data, a second training dataset was created. The modified dataset retains the same overall format as the original but updates the sentences to ensure that the target synsets differ while the context remains constant. To achieve this, lemmatization was applied to sentences to standardize word forms and ensure that the target word from the synset could be found within the sentence text. Consider the entry with ID `12070`: the verb “buzz” appears in the past tense as “buzzed” in the sentence, while the synset presents it in its base form “buzz”, without lemmatization the target word would not be correctly matched.

The general modification process works as follows: if the first synset is labeled more specific, I keep the second general sentence unchanged and modify the first sentence. I replace the target word in the general sentence with the more specific synset of the first sentence. Using the context of the more general sentence for both instances is advantageous because it is more likely to semantically suit both target words.

The overall reasoning behind this approach is that example sentences might not differ in specificity in the same way as the synsets they are illustrating. The specificity of other words and overall sentence length, can significantly impact a sentence's overall specificity. By keeping the context the same and changing only the target words, whose positions and relations in the semantic hierarchy are known, this method aims to provide a more controlled measure of specificity and maintain consistency in comparison, as illustrated in Table 1 ID 4.

When preparing the sentence pairs with the same context, several challenges were encountered. One significant issue arose from the difficulty of matching synset strings with sentences. Let's look at the instance with ID 238 in the Table 1: the synset 'clemency.n.02', extracted as a string "clemenency" was expected to be present in the sentence "He threw himself on the mercy of the court", but the term cannot be found even after lemmatization. This problem occurs because the synset's string represents the underlying concept and may not align precisely with the surface string in the sentences, especially when synonyms, paraphrasing, or idiomatic expressions are used.

Another challenge can be observed in the example with ID 2146. Here, the synsets capture related but distinct concepts of "luck" and "bad luck". However, both example sentences contain the concrete phrase "bad luck", which makes their specificity almost identical when placed in the same context. Such instances compromise the overall quality of the dataset and, consequently, the model's performance.

Additionally, modifying the sentences introduced grammatical issues. While lemmatization was necessary for aligning the data, it sometimes resulted in ungrammatical sentences, which had to be filtered out. For instance, in the entry with ID 16723 in Table 1, the target word was lemmatized to a base form that does not match the grammatical context of the sentence. Specifically, the lemma replacing the original target word was not conjugated to agree with the pronoun "she" in its third-person singular form.

After removing sentences where the target words could not be aligned and those with grammatical errors, about half of the initial dataset remains – around 8,500 data points. It is important to acknowledge that this substantial reduction in dataset size may further complicate the ability to draw clear comparisons regarding the effects of different training data.

3.3 Synthetic Data for Human Evaluation

To evaluate the performance of the specificity metric, human annotation was conducted. As outlined in Chapter 1, there is currently no standardized method for evaluating specificity metrics, which makes human judgment the most reliable approach for this purpose. In the experimental setup, each annotator was presented with 10 pairs of short article summaries. For each pair, annotators were required to select the summary they deemed more specific, or choose the “cannot decide” option if both summaries appeared equally specific and the choice was unclear. Each summary was assigned a specificity score based on the average score of all sentences within the text, normalized across the entire corpus of summaries. Subsequently, the binary choices made by the human annotators were compared with the model’s specificity scores to determine any correlation. This subsection will detail the process of generating the summaries used for human evaluation and discuss the characteristics of the data.

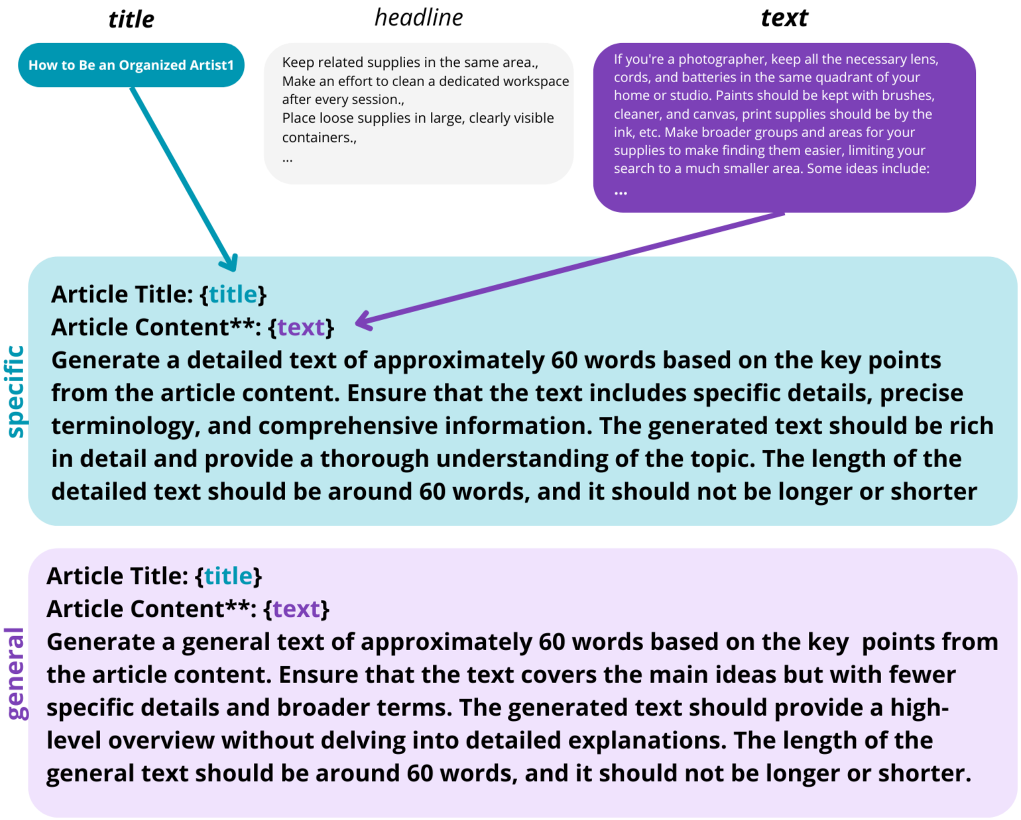
Although the metric was designed to measure specificity at the sentence level, I chose not to conduct human annotation on individual sentences. Judging a sentence’s specificity without context may be challenging for annotators, especially when they are untrained. Developing a consistent and fine-grained specificity scale that yields high inter-annotator agreement is also difficult. The only prior research using an ordinal scale for specificity by Gao et al. (2019) tailored their scheme to a specific data type and relied on crowdsourced annotations via Amazon Mechanical Turk. Unfortunately, the scope of this study did not allow for such large-scale resources.

Given these challenges, I decided to work with short texts rather than isolated sentences. However, asking people to assess a sentence or text in isolation can still be difficult. To simplify the task, I presented annotators with pairs of texts instead and asked them to choose which one appeared more specific. This comparative method made the task more intuitive and enabled clearer judgments.

The annotation task was undertaken by students with backgrounds in computational linguistics and other fields. All participants had English proficiency at a B2 level or higher, though their language skills varied. To reduce the potential impact of these variations, I selected the WikiSum dataset from Cohen et al. (2021) as a base for annotation. WikiSum consists of article summaries from the WikiHow website, featuring “how-to” articles and coherent paragraph summaries written in simplified English. [1]

[1] Simplified English or “Simple English” refers to a version of English that uses straightforward vocabulary and sentence structures to ensure clarity and accessibility for a broader audience, including non-native speakers and individuals with varying levels of literacy.

When preparing the data, I sampled 100 entries from the WikiSum dataset and extracted the titles and contents of the articles, with an average summary length of 676 words. Initially, I attempted to generate texts based solely on article headlines, but this approach resulted in summaries that were often dry and less coherent, as language models generally perform better when summarizing extensive content. Consequently, I focused on generating coherent summaries derived from the full article contents (see Figure 1).



Keep related supplies in the same area.,
Make an effort to clean a dedicated workspace after every session.,
Place loose supplies in large, clearly visible containers.,
...

If you're a photographer, keep all the necessary lens, cords, and batteries in the same quadrant of your home or studio. Paints should be kept with brushes, cleaner, and canvas, print supplies should be by the ink, etc. Make broader groups and areas for your supplies to make finding them easier, limiting your search to a much smaller area. Some ideas include:
...

Article Title: {title}
Article Content: {text}**
Generate a detailed text of approximately 60 words based on the key points from the article content. Ensure that the text includes specific details, precise terminology, and comprehensive information. The generated text should be rich in detail and provide a thorough understanding of the topic. The length of the detailed text should be around 60 words, and it should not be longer or shorter

Article Title: {title}
Article Content: {text}**
Generate a general text of approximately 60 words based on the key points from the article content. Ensure that the text covers the main ideas but with fewer specific details and broader terms. The generated text should provide a high-level overview without delving into detailed explanations. The length of the general text should be around 60 words, and it should not be longer or shorter.

Figure 1: Example prompts for generating specific vs. general summaries

I experimented with three different large language models: GPT-2, LLaMA 3.1, and Gemma2. GPT-2 did not produce comprehensive texts and mainly replicated the article's content, so it was excluded from further experiments.

For LLaMA 3.1, I tested various prompts, with the initial prompt shown in Figure 1. I set a 60-word limit to ensure summaries were brief yet detailed for evaluating specificity. Despite this limitation, the model frequently generated longer summaries for specific content, which could negatively influence the quality of human annotations due to discrepancies in length between pairs.

The results showed that the average specificity score for the general summary was slightly higher than that for the more specific summary. The more specific summary was often rated lower in terms of specificity, which contradicted initial expectations. The closeness of both scores to the midpoint of the specificity scale suggested that the averaging process might obscure intended distinctions in specificity. Furthermore, statistical tests revealed no significant differences between the distributions of the two summary types. Additionally, the more specific summaries often lacked coherence and comprehensiveness, which may have contributed to the unexpected results.

To refine the dataset for human annotation, I examined the distribution of differences between the two summary types. The density plot of these differences resembled a normal distribution, indicating that most differences were minimal (see Figure 2). Asking annotators to choose between pairs of texts that were similarly specific would likely yield limited insights, prompting me to filter out text pairs with minor differences in scores. Specifically, all summary pairs where the difference in specificity scores was within the range of -0.1 to 0.1 were removed. After this filtering process, 50 text pairs remained, ensuring that the pairs selected for annotation represented clearer differences in specificity.

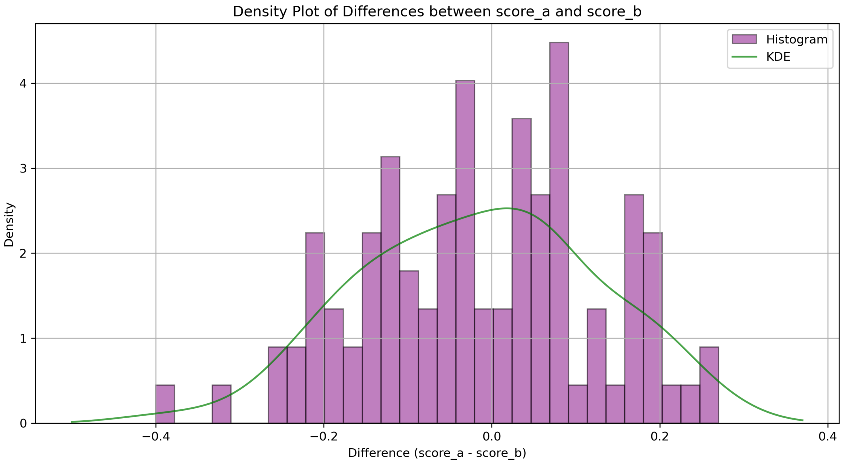


Figure 2: Density plot of difference between score_a and score_b

To ensure consistency in text length, I modified the prompt to check if responses fell within the 50-70 word range, truncating longer ones and regenerating shorter ones. I also adjusted the prompt to emphasize maintaining a natural flow while keeping a high level of detail. However, the issue of length persisted, as specific responses often required truncation, potentially making general texts less informative. Despite these changes, there was no significant difference in specificity scores between general and specific summaries.

As a next step, I decided to experiment with the temperature of the model, alongside the top_p parameter. The LLM temperature influences the language model's output by adjusting its randomness and creativity: a higher temperature results in more diverse and creative outputs, while a lower temperature yields more predictable text. The top_p parameter, also known as nucleus sampling, controls the diversity by considering only the smallest set of tokens whose cumulative probability exceeds a specified threshold. A higher top_p allows for more variety in the generated text, whereas a lower top_p focuses on more probable tokens.

For the experiments, I hypothesized that texts generated with higher temperature and top_p settings would be less specific, as the model might produce more varied and creative content that diverges from the exact instructions. The parameters for generating responses were set as follows: the specific prompt used a temperature of 0.7 and a top_p of 0.9, while the general prompt employed a temperature of 0.8 and a top_p of 0.95. Despite high hopes, these adjustments did not lead to any noticeable improvement in the distribution of specificity scores for the generated summaries.

In addition to refining the dataset, I explored alternative measures to capture text specificity more effectively. These measures included the maximum and minimum sentence-specificity values, range, standard deviation, and median of the specificity scores within a text. For most metrics, the difference between the distributions of `score_a` and `score_b` was not statistically significant. However, using standard deviation revealed a significant difference, likely due to the higher variability in texts generated with adjusted temperature and top_p settings, which increased the diversity of sentence specificity scores. Nonetheless, it was not suitable as a primary measure of specificity, as it reflects fluctuations in detail rather than the overall level of detail in the text.

I also experimented with the gemma2 model, using the same settings. However, the results worsened slightly, with `score_a` only surpassing `score_b` about half of the time. Additionally, the model often generated introductory phrases like "Here is the summary of the article" which had to be manually removed. These phrases made the text sound more AI-generated and negatively impacted the perceived specificity of the summaries.

Given that both models, despite carefully adjusted prompts, produced similar specificity scores for the general and specific summaries, it appears that the issue lies not with the language models themselves but with the specificity metric used in this evaluation. The fact that the current model struggles to consistently measure differences in specificity points to potential limitations in its design. I will discuss these issues further in the Chapter 5.

4 Model

In this section, we will explore the architecture of the model used to evaluate sentence specificity. The model utilizes a Siamese network framework built on BERT (Devlin et al., 2019), well-suited for comparing pairs of sentences. As illustrated in Figure 3, the model processes two sentences in parallel: `sentence1` and `sentence2`. For each sentence, BERT encodes the input to produce contextual embeddings for each token. These embeddings are fine-tuned during training. To obtain a fixed-size representation for each sentence, the embeddings are pooled by taking the mean of the token representations (mean pooling). These pooled embeddings are then processed through linear layers, ReLU activations, and dropout for regularization.The code and data used in this model can be found on GitHub. [2]

The Siamese network architecture computes logits for both sentences. To determine their relative specificity, the model calculates the difference between these logits. If the logit for `sentence2` is higher than the logit for `sentence1`, the difference will be positive and close to 1, indicating that `sentence2` is more specific. Conversely, if `sentence1` has a higher logit, the difference will be close to 0, reflecting that `sentence1` is more specific. This difference is transformed into a probability score through a sigmoid function, ensuring that the output is between 0 and 1. While the model outputs binary labels during training to indicate which sentence is more specific, the primary goal is to generate continuous specificity scores for both sentences.

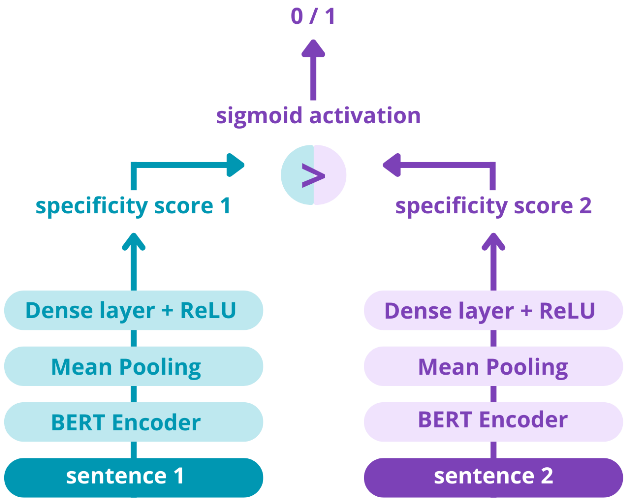
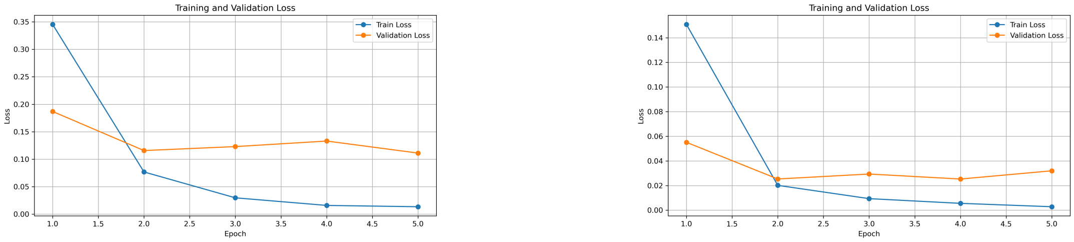


Figure 3: Model architecture

[2] `https://github.com/ydarja/BA-thesis`



(a) Trained on raw sentences

(b) Trained on SC sentences

Figure 4: Training and validation loss

Both models, trained on raw sentences and on sentences with the same context, shared the same set of optimal hyperparameters. A smaller batch size of 4 was found to significantly improve accuracy. The optimal number of epochs was determined to be 2, as indicated in Figure 4, which shows that after the second epoch, the model begins to overfit the training data as validation loss increases. The learning rate was set to 1×10^{-5}, and a dropout rate of 0.3 was employed to prevent overfitting.

5 Results

5.1 Statistical Analysis of the Specificity Scores

This section presents the statistical analysis of the specificity scores computed for the dataset and their correlation with two key factors: WordNet depth and sentence length. The results discussed in this section focus on the model trained on raw example sentences, as it demonstrated stronger performance compared to the model trained on sentences with controlled context. WordNet depth and sentence length are considered indicative of the level of detail and semantic complexity in text, and analyzing their relationship with specificity can provide insights into the efficiency of the proposed metric.

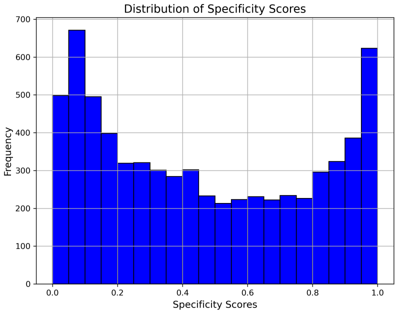


Figure 5: Specificity score distribution on the test split of raw sentence dataset

The descriptive statistics for specificity scores, WordNet depths, and sentence lengths are summarized in Table 2. The mean specificity score is 0.467, with a median of 0.418 and a standard deviation of 0.325. This suggests a moderate average specificity across the dataset, with considerable variation.

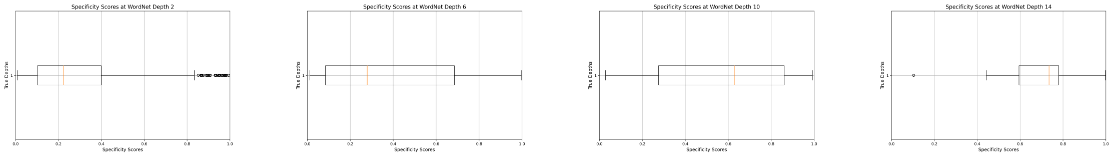
Statistical Metric	**Specificity Score**	**WordNet Depth**	**Sentence Length**
Mean	0.471	5.734	6.931
Median	0.417	6.000	7.000
Standard Deviation	0.325	2.812	2.875

Table 2: Statistics of Specificity Score, WordNet Depth, and Sentence Length

The U-shaped distribution of specificity scores, shown in Figure 5, indicates a polarization where many sentences are either highly specific or very general, with fewer falling in between. This polarization could be related to the nature of the model's training, where it is explicitly tasked with choosing the more specific of two sentences. As a result, the model may exaggerate the difference between general and specific sentences.

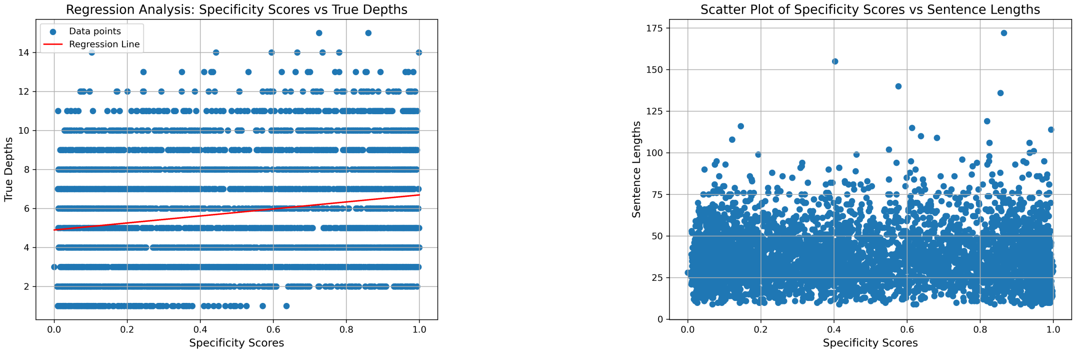
The average WordNet depth of 5.73 suggests the dataset contains a mix of general and specific terms, contributing to this variation in specificity. Similarly, the average sentence length of approximately 7 words, with noticeable variability, indicates a tendency toward shorter sentences. This could be influenced by the presence of shorter noun and verb phrases in the dataset.

The boxplots of specificity scores at various WordNet depths are illustrated in Figure 6. At shallower depths, the distribution is skewed to the left, indicating a tendency toward more general terms. As the depth increases, the distribution shifts and becomes increasingly skewed to the right. This trend indicates a relationship between WordNet depth and specificity, suggesting that deeper terms tend to yield more specific vocabulary.



(a) depth=2 (b) depth=6 (c) depth=10 (d) depth=14

Figure 6: Specificity scores at various WordNet depths



(a) Specificity scores vs WordNet depths

(b) Specificity scores vs sentence length

Figure 7: Correlation analysis of specificity scores

The correlation analysis offers some insights regarding specificity scores in relation to WordNet depths and sentence lengths. The Pearson correlation coefficient between specificity scores and WordNet depths is approximately 0.208, while the Spearman correlation is about 0.196. This indicates a modest positive correlation, which aligns with expectations, as the model's training process inherently links specificity to the depth of the terms utilized. The regression plot in Figure 7a further illustrates a slight upward trend for specificity scores versus WordNet depths. The regression analysis reveals an R-squared value of 0.043.

In contrast, the correlation coefficients for specificity scores and sentence lengths are quite low, with a Pearson correlation of approximately 0.029 and a Spearman correlation of 0.014. The scatter plot for specificity scores against sentence lengths in Figure 7b also shows no real relationship between specificity scores and length of the sentences. This lack of correlation suggests that sentence length does not serve as a reliable indicator of specificity in this dataset. However, such results are not surprising, as the model did not take length of the sentences into account during training.

5.2 Comparison with Other Specificity Metrics

To further analyze the created specificity metric, I examined its correlation with another openly available metric from the MoreThanSentiments package (Jiang and Srinivasan, 2023). They narrowly define specificity as a measure of the quality of relating uniquely to a particular subject. The specificity metric is calculated as the number of entity names, quantitative values, and times/dates scaled by the total number of words in a document.

When comparing this metric to mine, the Pearson correlation coefficient is 0.0343, indicating no correlation. As shown in the scatter plot (Figure 8), there is no trend between the two metrics. This result is not entirely surprising, given the differing definitions of specificity, which limits the value of direct comparison. Additionally, my metric is designed for sentence-level measurement, while theirs is applied at the document level.

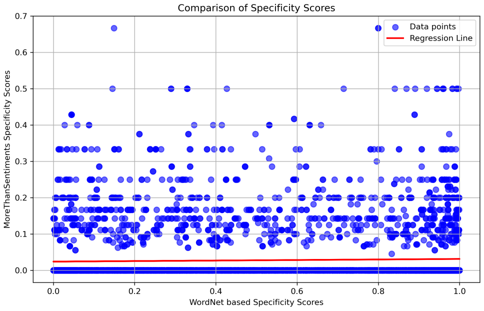


Figure 8: Comparison with MoreThanSentiments specificity scores

To further validate the proposed specificity metric, I compared it with the other existing metric trained on Twitter data (Gao et al., 2019). Their model was trained on a dataset of over 7,000 tweets annotated on a fine-grained specificity scale from 1 to 5. Since I encountered difficulties running their model on my test data, I opted to use their human-annotated test dataset of 700 tweets for comparison. I preprocessed these tweets by cleaning them of any links, emojis and user mentions. Links and user mentions were removed as they were masked in the original dataset, while emojis were excluded to maintain a focused analysis on the textual content. . I also removed any tweets longer than a single sentence as my metric is designed for sentences, not texts. The original ordinal scores were normalized to a continuous scale. This allowed for a meaningful analysis of how the two metrics correlate when applied to similar data.

After normalizing the Twitter specificity scores and applying my own model to generate scores for the same dataset, I calculated the Pearson correlation between the two sets of scores, which turned out to be fairly high at 0.38. The trend is also clearly seen in the plot in Figure 9.

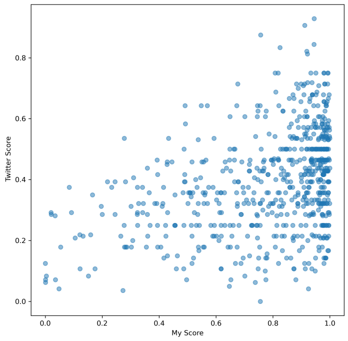
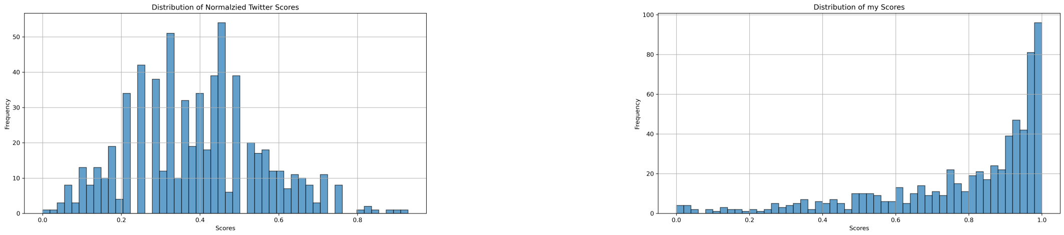


Figure 9: Correlation of my scores with Twitter scores

However, when comparing the statistics of my scores with the Twitter scores, significant differences emerge (see Figure 10). The distribution of Twitter specificity scores is bell-shaped, suggesting that most tweets exhibit mid-level specificity, mixing general and specific information. In contrast, my model's scores are skewed to the right, indicating that most tweets received high specificity values.

The statistics confirm this: the mean of my scores is 0.78 with a standard deviation of 0.23, while the normalized Twitter scores have a mean of 0.39 and a standard deviation of 0.17. This discrepancy likely arises due to differences in the data and criteria used by the models. The Twitter data associates specificity with mentions of concrete entities, such as named persons or events, whereas my model was trained on WordNet and focuses on abstract versus specific word meanings. Additionally, I assume that the vocabulary of tweets differs significantly from the WordNet training data, introducing many novel words to my model, which may further contribute to the inflated specificity scores. The broader, more informal language in social media posts, combined with the limited lexical scope of my training data, could amplify this effect.



(a) Twitter specificity scores

(b) My specificity scores

Figure 10: Distributions of my scores vs Twitter-based specificity scores

5.3 Human Evaluation

As described in Chapter 3.3, for human evaluation I selected 50 pairs of specific and general summaries generated from WikiHow articles using the LLaMA 3.1 model. The prompt consistently instructed the model to produce the first summary as more specific and the second as more general. However, when I analyzed the specificity scores generated by my model for these summaries, I found no significant difference in their distributions. This led to an even distribution of gold labels regarding which summary was deemed more specific. Specifically, if the first summary was more specific, it was assigned a label of 0; if the second summary was more specific, it received a label of 1.

The entire dataset was divided into five batches, each containing ten data points. Each batch was annotated by three annotators: two students from computational linguistics and one student from an unrelated program. The responses were fully anonymized. Participants were required to have at least a B2 proficiency level in English to take part in the study. The survey took approximately 10 to 15 minutes to complete, which helped ensure a high level of concentration during the annotation process.

When analyzing the overall distribution of responses in Figure 11, we find that 61.4% of choices favored the first summary, while the remaining responses were evenly divided between the second summary and the “cannot decide” option. These results are noteworthy, as they suggest that human annotators largely align with the LLM’s interpretation of specificity in summarization. This agreement raises questions about the effectiveness of my model in capturing human perceptions of summary quality.

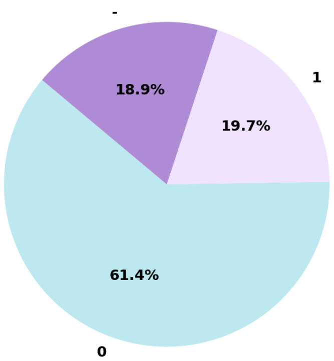


Figure 11: Distribution of human annotation choices (0,1,-)

The inter-annotator agreement analysis, summarized in Table 3 reveals varied Kappa scores across batches. Batches 1 and 2 indicate fair agreement, while Batches 3, 4 and 5 reflect poor agreement, resulting in a low average Kappa score of 0.07. Several factors may explain the inconsistent results: annotators might have struggled to distinguish between “more specific” and more “informative”, leading to confusion in their judgments. Some participant reported that they found that some summary pairs semantically conveyed the same information with different wordings, further complicating evaluations. Additionally, batches with lower agreement may have included summaries with smaller average differences in specificity.

Batch	Mean Kappa Score
Batch 1	0.28
Batch 2	0.32
Batch 3	-0.15
Batch 4	0.05
Batch 5	-0.11
Average	0.07

Table 3: Mean Kappa scores for each batch

The correlation analysis between human annotators' choices and the model's specificity scores was conducted using point-biserial correlation, yielding a coefficient of 0.19 with a p-value of 0.05. In this analysis, the difference between the model's specificity scores for two summaries, calculated as `score_b - score_a`, was compared against the binary human labels. A negative difference indicates that summary A is more specific than summary B, corresponding to a human judgment of 0. Conversely, a positive difference suggests that summary B is more specific, which aligns with a human judgment of 1.

While the observed correlation aligns with expectations, the p-value of 0.05 indicates marginal statistical significance. This suggests that the model's specificity scores capture some structure related to human judgments, but the realtion remains weak. It's important to note that the dataset used for this analysis is quite small, which may limit the reliability of the findings. These results underscore the need for further refinement in the model's training and a clearer definition of specificity to improve alignment with human evaluations.

6 Discussion and Conclusions

In this work, we introduced a new specificity metric based on WordNet and evaluated its performance. While the results demonstrated some promise, there are several areas for future improvements.

The model was trained on sentence pairs from WordNet and effectively generated specificity scores, revealing a U-shaped distribution. It indicates a polarization of sentences into highly specific or very general categories. This pattern likely stems from the nature of the training process, where the model is tasked with selecting the more specific sentence in each pair. As a result, it may exaggerate the specificity differences, making it harder to capture more nuanced, intermediate cases.

The moderate Pearson correlation of 0.38 between the model's specificity scores and human-labeled data from Twitter reflects a positive outcome of this study. The model aligns well with human judgments despite the differences between the tweet data and the training data. This result is noteworthy given that the Twitter test dataset consists of 700 sentences, significantly larger than my own smaller-scale annotation, and was annotated by professional annotators, ensuring the reliability of the gold scores.

However, in my own human evaluation, the correlation analysis revealed a coefficient of 0.179 (p-value 0.05), indicating that while the model's specificity scores capture some structure, they do not consistently align with human judgments. Additionally, the overall agreement between the model predictions and my human assessments was low, with an average Kappa score of 0.07. This suggests a potential misalignment between the model's scoring mechanism and human intuition, likely due to the limited data available and subjective nature of human evaluations.

The analysis also showed a positive correlation between specificity and WordNet depth. While this outcome reflects the design of the training process, where synset depth influenced the selection of more specific sentences, it still reinforces the utility of WordNet depth as a meaningful measure of specificity. However, the lack of significant correlation between sentence length and specificity, with a Pearson correlation of 0.029, indicates a potential limitation in the current metric. It is not entirely surprising, as sentence length was not explicitly considered during training. While longer sentences are often assumed to be more specific, the model's inability to capture this highlights the need for future refinements, including incorporating sentence length as one of the factors in specificity predictions.

These findings suggest that the current approach requires substantial refinement to fully align with human evaluations and adequately capture the nuances of text specificity. To improve the accuracy of specificity assessments, it is essential to consider the limitations inherent in the current methodology and outline directions for future work.

One of the main limitations of the current metric is its reliance solely on WordNet depth, which fails to define specificity on its own. A more robust metric could be developed by incorporating additional factors beyond WordNet depth. For instance, taking inspiration from previous research, the model could be trained on key syntactic and semantic features described in Chapter 2.1, such as phrase structure, grammatical roles, and semantic relationships. In addition, leveraging word embeddings (e.g., BERT or GPT) could enable a more context-aware approach, where not just the hierarchical depth of a word is considered, but also the richness of its meaning within a given sentence. By compressing or extracting relevant semantic information from embeddings, the specificity score could reflect not only lexical choice but also syntactic complexity and sentence structure.

The training dataset would also benefit from some improvements. In the current setup, when we train the model on sentence pairs with the same context, some pairs of sentences might become nonsensical when replacing target words, which could confuse both the model. A more careful filtering process could be introduced by first feeding the sentences into a language model and calculating a perplexity or surprisal score. By setting a threshold for these scores, it would be possible to automatically remove sentences that are grammatically incorrect or that do not make sense in context. This would ensure that the model is trained and evaluated on cleaner, more coherent data. Such filtering was not possible during this study due to the already limited amount of training data.

In terms of human evaluation, there are several ways the process could be improved. One option is to move beyond binary choices and ask annotators to provide a specificity score on a scale. This would allow for more granular judgments and better alignment with the continuous nature of the metric. Additionally, other factors like coherence and readability could be considered to provide a fuller assessment of the summaries. To ensure the reliability of annotations, an obvious example could be introduced in the beginning of the survey to filter out participants who may not fully understand the task or who may not be paying adequate attention.

A more extensive human evaluation is also necessary to allow for more grounded conclusions. This study involved three annotators and five batches, expanding the pool of participants and including more data points would provide greater reliability and generalizability. Furthermore, comparing this metric with existing specificity measures in practical applications could help contextualize its effectiveness. For example, fine-tuning a summarization model using three different sets of reference summaries (randomly picked, based on my metric, and based on existing metrics) could serve as a practical test of whether the metric is truly useful in real-world applications. By comparing the performance of models fine-tuned with these different datasets, it would be possible to draw conclusions about the metric's value in NLP domain.

In addition to these improvements, an interesting direction for future research would be to conduct a psycholinguistic study focused on how specificity influences reader behavior. Although this falls outside the scope of the current study, it would offer valuable insights into how varying levels of specificity affect attention, cognitive load, and reading time. By understanding the cognitive processes involved when readers encounter more or less specific text, we could gain a deeper understanding of how specificity impacts the comprehension and retention of information. Furthermore, studies have shown connections between the use of specific language and cognitive processing in populations such as those with autism Li et al. (2017).

In summary, while the current metric offers an initial attempt to quantify specificity, there are clear opportunities for improvement. By incorporating more advanced linguistic features, enhancing dataset quality, refining human evaluation methods, and testing the metric in practical applications, it could be significantly improved. These improvements would not only enhance the reliability of the metric but also promote a clearer evaluation of large language models (LLMs) in natural language generation tasks. A strong specificity metric could be a valuable tool for assessing the quality of generated texts, ensuring they achieve the essential standards of informativeness and precision. By deepening our understanding of specificity in generated outputs, this metric has the potential to play a significant role in developing more effective NLG systems, ultimately improving communication in various automated contexts.

References

Cohen, N., O. Kalinsky, Y. Ziser, and A. Moschitti (2021, August). WikiSum: Coherent summarization dataset for efficient human-evaluation. In C. Zong, F. Xia, W. Li, and R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Online, pp. 212–219. Association for Computational Linguistics.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio (Eds.), *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)*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics.

Fabbri, A. R., W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev (2021). Summeval: Re-evaluating summarization evaluation.

Gao, Y., Y. Zhong, D. Preoţiuc-Pietro, and J. J. Li (2019, Jul.). Predicting and analyzing language specificity in social media posts. *Proceedings of the AAAI Conference on Artificial Intelligence 33*(01), 6415–6422.

Hu, X., M. Gao, S. Hu, Y. Zhang, Y. Chen, T. Xu, and X. Wan (2024). Are LLM-based evaluators confusing NLG quality criteria?

Jiang, J. and K. Srinivasan (2023). Morethansentiments: A text analysis package. *Software Impacts 15*, 100456.

Ko, W.-J., G. Durrett, and J. J. Li (2018). Domain agnostic real-valued specificity prediction.

Levshina, N. (2022, 02). Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy 24*, 280.

Lex, E., M. Völske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein, and M. Granitzer (2012, 04). Measuring the quality of web content using factual information. pp. 7–10.

Li, J. and A. Nenkova (2015, Feb.). Fast and accurate prediction of sentence specificity. *Proceedings of the AAAI Conference on Artificial Intelligence 29*(1).

Li, J. J., J. Parish-Morris, L. Bateman, and A. Nenkova (2017). Autism quotient scores modulate the perception and production of text specificity in adult females. In *International Meeting for Autism Research*.

Lin, C.-Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Barcelona, Spain, pp. 74–81. Association for Computational Linguistics.

Loper, E. and S. Bird (2002). NLTK: The natural language toolkit.

Louis, A. (2013, 01). Predicting text quality: Metrics for content, organization and reader interest.

Louis, A. and A. Nenkova (2012, 05). General versus specific sentences: automatic identification and application to analysis of news summaries. *Technical Reports (CIS)*.

Meng, L., R. Huang, and J. Gu (2014, 06). Measuring semantic similarity of word pairs using path and information content. *International Journal of Future Generation Communication and Networking 7*, 183–194.

Miller, G. A. and W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*(1), 1–28.

Navigli, R. (2009, 02). Word sense disambiguation: A survey. *ACM Comput. Surv. 41*.

Nguyen, H., H. Chen, L. Pobbathi, and J. Ding (2024). A comparative study of quality evaluation methods for text summarization.

Novikova, J., O. Dušek, A. Cercas Curry, and V. Rieser (2017, September). Why we need new evaluation metrics for NLG. In M. Palmer, R. Hwa, and S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2241–2252. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002, July). Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, pp. 311–318. Association for Computational Linguistics.

Raiaan, M., M. S. Hossain, K. Fatema, N. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam (2024, 01). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access PP*, 1–1.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy.

Richardson, R., A. F. Smeaton, and J. Murphy (1994). Using wordnet as a knowledge base for measuring semantic similarity between words.

Rudnicka, E., F. Bond, Ł. Grabowski, M. Piasecki, and T. Piotrowski (2018, January). Lexical perspective on Wordnet to Wordnet mapping. In F. Bond, P. Vossen, and C. Fellbaum (Eds.), *Proceedings of the 9th Global Wordnet Conference*, Nanyang Technological University (NTU), Singapore, pp. 209–218. Global Wordnet Association.

Yao, K., B. Peng, G. Zweig, and K.-F. Wong (2016). An attentional neural conversation model with improved specificity.

Zhang, R., J. Guo, Y. Fan, Y. Lan, J. Xu, and X. Cheng (2018, July). Learning to control the specificity in neural response generation. In I. Gurevych and Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, pp. 1108–1117. Association for Computational Linguistics.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi (2020). Bertscore: Evaluating text generation with bert.