Data Augmentation for Low-Resource Italian NLP: Enhancing Semantic Processing with DRS

Muhammad Saad Amin*, Luca Anselma and Alessandro Mazzei

Department of Computer Science, University of Turin, Italy

Abstract

Discourse Representation Structure (DRS), a formal meaning representation, has shown promising results in semantic parsing and natural language generation tasks for high-resource languages like English. This paper investigates enhancing the application of DRS to low-resource Italian Natural Language Processing (NLP), in both semantic parsing (Text-to-DRS) and natural language generation (DRS-to-Text). To address the scarcity of annotated corpora for Italian DRS, we propose a novel data augmentation technique that involves the use of external linguistic resources including: (i) WordNet for common nouns, adjectives, adverbs, and verbs; (ii) LLM-generated named entities for proper nouns; and (iii) rule-based algorithms for tense augmentation. This approach not only increases the quantity of training data but also introduces linguistic diversity, which is crucial for improving model performance and robustness. Using this augmented dataset, we developed neural semantic parser and generator models that demonstrated enhanced generalization ability compared to models trained on non-augmented data. We evaluated the effect of semantic data augmentation using two state-of-the-art transformer-based neural sequence-to-sequence models, i.e., byT5 and IT5. Our implementation shows promising results for Italian semantic processing. Data augmentation significantly increased the performance of semantic parsing from 76.10 to 90.56 (+14.46%) F1-SMATCH score and generation with 37.79 to 57.48 (+19.69%) BLEU, 30.83 to 40.95 (+10.12%) METEOR, 81.66 to 90.97 (+9.31%) COMET, 54.84 to 70.88 (+16.04%) chrF, and 88.86 to 92.97 (+4.11%) BERT scores. These results demonstrate the effectiveness of our novel augmentation approach in enhancing semantic processing capabilities for low-resource languages like Italian.

Keywords

Data augmentation, Italian semantic processing, low-resource NLP, semantic parsing and generation

1. Introduction

The field of Natural Language Processing (NLP) has seen significant advancements in recent years, particularly in semantic processing tasks. These tasks, which include semantic parsing and natural language generation, often rely heavily on parallel corpora — datasets that align text in one language with its semantic representation or with text in another language [1, 2]. For languages with rich linguistic resources, such as English, the availability of large-scale parallel corpora has facilitated rapid progress in semantic processing [3, 4]. However, for many languages, including Italian, the scarcity of such resources poses a significant challenge to advancing semantic NLP capabilities [5, 6]. Italian presents unique challenges and opportunities. While Italian shares some structural similarities with English, it possesses distinct linguistic features that complicate NLP tasks. These include a more flexible word order, a rich system of verb conjugations, and the presence of grammatical gender for nouns, adjectives, and articles.

In the context of NLP and Natural Language Generation (NLG), Italian has seen moderate progress. However, compared to high-resource languages like English, Italian still lacks extensive task-specific datasets, particularly in areas requiring deep semantic understanding. This deficiency is especially pronounced in tasks involving formal semantic representations such as Discourse Representation Structures (DRS) [7].

While Italian is not typically classified as a low-resource language in general NLP terms, it can be considered as such in the specific domain of semantic processing, especially when dealing with formal semantic representations. This status is characterized by: (i) Named Entities: Italian naming conventions differ from those in English, requiring adaptation in entity recognition tasks; (ii) Syntactic Structure: Although Italian follows the SVO structure like English, it allows for greater flexibility, posing challenges, especially in parsing tasks; (iii) Grammatical Gender: The presence of grammatical gender in Italian adds complexity to tasks such as coreference resolution and agreement in the generated text. These linguistic features, combined with the limited availability of semantically annotated corpora, position Italian as a challenging language for advanced semantic NLP tasks.

Data augmentation (DA), a technique widely used in machine learning to increase the size and diversity of training datasets, has shown promise in addressing re-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

muhammadsaad.amin@unito.it (M. S. Amin); luca.anselma@unito.it (L. Anselma); alessandro.mazzei@unito.it (A. Mazzei)

0000-0002-7002-9373 (M. S. Amin); 0000-0003-2292-6480 (L. Anselma); 0000-0003-3072-0108 (A. Mazzei)

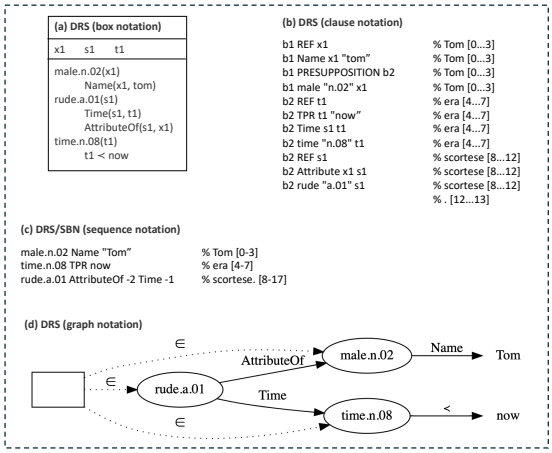

Figure 1: Different graphical representations of DRS for the text "Tom era scortese." or "Tom was rude."

source scarcity in NLP [8]. For semantic tasks involving DRS, DA presents unique challenges due to the need to preserve semantic equivalence while introducing linguistic variety.

In the context of Italian semantic processing, traditional augmentation techniques such as random word insertion, deletion, substitutions or back-translation have limited applicability due to the scarcity of Italian-specific semantic resources [9]. This necessitates innovative approaches that can leverage resources from high-resource languages while maintaining the integrity of Italian linguistic structures.

Given the challenges outlined, this study aims to develop a novel cross-lingual DA technique for Italian, specifically tailored for DRS-based semantic parsing and generation tasks. While word substitution techniques are established in DA literature, our approach introduces an innovative cross-lingual framework that leverages the language-neutral nature of DRS. The method uniquely bridges the resource gap between high-resource and low-resource languages by temporarily transforming Italian examples into English, enabling access to rich lexical resources like WordNet, before converting back to Italian. This cross-lingual approach leverages the universal semantic representations of the DRS to enable more advanced data transformation approaches than Italian resources alone would allow, which is particularly advantageous given the limited availability of Italian-specific semantic datasets (see Table 1 for Italian examples).

This paper makes the following key contributions:

1. A novel cross-lingual augmentation methodology that leverages English WordNet to enhance Italian semantic datasets.
2. Empirical evidence demonstrating the effectiveness of this augmentation technique in improving performance scores for both DRS parsing and generation tasks in Italian.
3. A detailed analysis of how cross-lingual augmentation affects the handling of Italian-specific linguistic features in semantic processing.
4. Insights into the scalability and potential applications of this approach to other low-resource languages in the domain of semantic NLP.

The remaining paper is organized as follows: Section 2 provides an overview of DRS. Section 3 details semantic DA for Italian with a focus on named entities, lexical, and grammatical data transformation techniques. Section 4 presents our experimental implementation, implications of our results and findings, and their broader impact on the field. Finally, Section 5 concludes the paper, addresses certain limitations, and outlines directions for future research.

2. Background

In this Section, we provide an overview of the formal definition of DRS.

DRS is a formal semantic representation, that captures the essential meaning of text, equivalent to first-order logic. DRS is capable of representing a broad spectrum of linguistic phenomena, including anaphora, presuppositions, and temporal expressions [7]. What sets DRS apart from other meaning representations, such as Abstract Meaning Representation (AMR) [2], is its proficiency in handling negation and quantification, as well as its language-independent nature. Furthermore, DRS can effectively represent meaning across multiple sentences in a discourse.

Initially, DRS utilized box notation to provide scope to meaning representation (see Figure 1(a)). This notation incorporates (e.g. *x1*) and conditions (e.g. *person*, *Time*), with concepts anchored using WordNet synsets and thematic roles derived from VerbNet. Operators (e.g. $=$) are employed to establish comparative relationships between entities. Conditions can also embody complex structures to express logical (e.g. NEGATION, $\neg$) or rhetorical relationships among various condition sets. To address the challenges posed by the complexity of box notation in neural parser development, Clause Notation was introduced. This method streamlines DRS by reorganizing the structure and placing variables before discourse referents and conditions (see Figure 1(b)).

Further simplification led to the development of Sequence Box Notation (SBN), a variable-free format designed to be more compatible with neural sequence-to-sequence transformer architectures [7]. SBN utilizes indices to form connections between concepts, with thematic roles indicating the nature of these connections (see Figure 1(c)). This notation can also be interpreted in

graph form (see Figure 1(d)). These evolving notations reflect the ongoing efforts to make DRS more accessible and efficient for computational processing while maintaining its rich semantic representation capabilities.

3. Semantic DA for Italian

The data-intensive nature of neural networks presents a significant challenge for low-resource languages like Italian, where available data is limited. This challenge is further compounded when dealing with logical semantic representations such as DRS-Text pairs, which follow specific patterns. In DRS, concepts are represented as a combination of lemma, part of speech, and WordNet sense numbers. The part of speech component includes adjectives, adverbs, common nouns, and verbs with lexical entities, followed by other logical representations (e.g., "idea.n.01").

Our augmentation methodology addresses the scarcity of Italian lexical resources by utilizing a cross-lingual approach that takes advantage of the language-neutral structure of DRS. The process (i) begins with translating the Italian text into English while keeping the original DRS unchanged; (ii) allowing us to apply a variety of augmentation techniques including named-entity, lexical, and grammatical augmentations—made possible through access to English WordNet—on English-aligned examples; (iii) after augmentation, the English examples are translated back into Italian, ensuring that the semantic relationships from the DRS are preserved. This strategy not only generates semantically rich and contextually relevant data but also overcomes the limitations of Italian-specific resources by augmenting English-aligned examples and transforming them into Italian-aligned examples (see Figure 2 and Table 4 in Appendix), maintaining semantic accuracy through DRS's formal representations.

3.1. Named Entities Augmentation

Our initial augmentation approach focused on proper noun (PN) augmentation, also referred to as Named Entities (NE) Augmentation. This method targets the transformation of specific named entities, particularly person names (PER, both male and female) and geographical entities (GPE) such as city, state, country, and island names. These entities are explicitly represented in the DRS through predicates (e.g., "male.n.02" for person names). We employed a rule-based approach to extract NEs from both the DRS and the text. Our NE augmentation strategy involves replacing existing entities with those outside the context of the dataset. This approach aims to evaluate the role of external lexical information in semantic processing.

To maintain semantic integrity, we ensure that NEs are replaced with entities of the same type. For sourcing external lexical information, we utilized AI-generated lists of person names based on global frequency and GPE entities with similar geographical distribution, carefully filtering out names already present in the dataset. This meticulous substitution process preserves the true semantics of the sentences. For instance, in the sentence "Rome is the capital of Italy", we might replace "Rome" with "Berlin" and "Italy" with "Germany", maintaining the logical structure while introducing lexical variety.

3.2. Lexical Entities Augmentation

Our lexical augmentation strategy focuses on four specific categories: common nouns, adjectives, adverbs, and verbs. We utilize WordNet synsets to group these entities, ensuring that transformations maintain the contextual sense and meaning of the sentences.

Common Noun Augmentation: CN can significantly alter sentence meaning, making their augmentation challenging. We employ a rule-based approach to extract common nouns from the Sequence Box Notation (SBN) and use NLTK's "WordNetLemmatizer" for the corresponding text. The augmentation process involves replacing nouns with their hyponyms from WordNet, which allows for more specific substitutions while preserving contextual meaning.

Verb Augmentation: Verbs play a crucial role in sentence context, making their augmentation complex. We use WordNet-based troponyms to replace verbs with more specific, contextually similar alternatives. This approach helps maintain semantic coherence while introducing lexical variety.

Adjective Augmentation: Adjectives, as descriptive attributes of nouns, are augmented using WordNet-based antonyms. This method generates new, contextually similar examples. We manually inspect the augmented data to ensure the semantic relevance and correctness of adjective substitutions.

Adverb Augmentation: For adverbs, we employ a WordNet-based synonym replacement approach. This method aims to generate similar data examples while preserving contextual relevance. As with other categories, we manually verify the semantic correctness of the newly generated examples. Throughout the augmentation process for all lexical categories, we maintain consistency between the SBN logical representations and the corresponding text. This ensures that the augmented data remains coherent and semantically valid across both the formal representation and natural language formats.

3.3. Grammatical Augmentation

This approach primarily focuses on transforming morpho-syntactic relations within sentences, with a par-

ticular emphasis on tense modifications. This method involves non-lexical substitutions that alter the temporal context of events without introducing external vocabulary. Our strategy encompasses a wide range of grammatical transformations, including shifts between present, past, and future tenses, as well as changes in voice (active to passive and vice versa), mood (e.g., imperative), negation, number (singular to plural), subject-object relationships, aspect (progressive and perfect), and other grammatical features such as infinitive forms, first-person perspective, and perfect participles.

To implement these transformations, we employ a dual approach: for the Sequence Box Notation (SBN), we use a rule-based system to replace logical entities (e.g., changing "EQU" to "TPR" or "TSU" for tense shifts), while for the corresponding natural language text, we utilize the *tenseflow API*[1]. This comprehensive grammatical augmentation technique allows us to significantly expand our dataset with grammatically diverse versions of existing sentences, maintaining core semantic content while introducing new syntactic variety. Such diversity is essential for training robust NLP models, particularly for tasks involving temporal reasoning and varied syntactic structures.

While our augmentation strategies effectively expand the dataset nine times, we acknowledge specific challenges in preserving semantic integrity during transformations. For named entities, semantic preservation is straightforward as we maintain entity types. However, tense transformations present more complexity due to Italian's rich verbal morphology. For instance, the Italian imperfetto tense ("cantava"–was singing) can map to multiple English past tense forms, requiring careful handling to maintain the original temporal relations in the DRS. Additionally, Italian's pro-drop nature and flexible word order can complicate the preservation of argument structure when performing verbal augmentations.

4. Experimental Implementation

Our experimental setup utilizes the Italian, German, Dutch, and English versions of logic-text pairs from the Parallel Meaning Bank (PMB) release 5.0.0[2] [10] (statistical numbers for multilingual baselines are listed in Table 1). These datasets are categorized into three annotation levels: Gold (fully manually annotated), Silver (partially manually annotated), and Copper (machine-translated version of English data examples without any annotation). For Italian meaning representation, we maintain this annotation distinction. We adhere to the same data split for training, development, and test sets [10]. Each data example consists of a pair: a DRS meaning representation and its corresponding textual form.

Table 1
Dataset split along with statistic numbers for multi-lingual baselines. Note: T_Gold = Train Gold; T_Silver = Train Silver

Langs	T_Gold	Dev	Test	T_Silver
Italian	745	555	555	4,316
German	1,206	900	900	6,862
Dutch	586	435	435	1,646
English	9,057	1,132	1,132	143,731

Categorization of Augmented Data: To facilitate a comprehensive analysis of our augmentation strategies, we classify the augmented dataset into various categories based on named entities, lexical, and grammatical transformations. Our experimental approach is structured into three main categories: (i) baseline experiments without augmentation; (ii) individual augmentation — applying one augmentation technique at a time; and (iii) compound augmentation — concatenating all augmentation approaches applied to the Italian semantic corpus. Table 2 provides detailed information on the types of augmentation, dataset sizes, and the number of training examples for both individual and compound augmentation strategies employed in our experiments.

Table 2
Impact on the size of Italian dataset examples without augmentation and with individual and compound augmentation. Note: w/o = without; Aug = Augmentation; Ex. = Examples; G = Gold; S = Silver; G-S = Gold-Silver; CN = Common Noun; NE = Named Entities; Adj. = Adjectives; Adv = Adverbs; Comp = Compound

Training Type	Size	# G Ex.	# S Ex.	# G-S Ex.
w/o Aug	x1	745	4316	5061
NE Aug	x2	1490	8632	10122
CN Aug	x2	1490	8632	10122
Adj Aug	x2	1490	8632	10122
Adv Aug	x2	1490	8632	10122
Verb Aug	x2	1490	8632	10122
Tense Aug	x4	2980	17264	20244
Comp Aug	x9	6705	38844	45549
Dev	–	555	–	–
Test	–	555	–	–

Neural Architecture Our approach to semantic parsing and generation primarily involves fine-tuning the byT5 model [11], a multilingual variant of the T5 transformer. We chose byT5 for several compelling reasons: (i) its multilingual nature enhances cross-language and cross-task generalization; (ii) its byte-level tokenization

[1] https://github.com/bendichter/tenseflow

[2] The PMB is developed at the University of Groningen as part of the NWO-VICI project "Lost in Translation – Found in Meaning" (Project number 277-89-003), led by Johan Bos.

Table 3
Italian semantic parsing and generation results of byT5 and IT5 with multi-lingual baselines and augmentation on PMB-5.0.0. The best results are **bold** and underlined. (Aug = Augmentation; Adj = Adjective; Adv = Adverb; NE = Named Entities; CN = Common Noun; Comp = Compound; G = Gold; S = Silver; C = Copper).

Exp.	Impl. Type	Dataset	Parsing Results	Generation Results				
		Flavour	SMATCH (F1%)	BLEU	BERT-Score	METEOR	COMET	chrF
1	German	G+S	73.00	34.14	88.24	30.07	59.53	53.72
2	Dutch	G+S	42.77	19.83	84.98	25.36	51.78	46.92
3	English	G+S	91.42	71.89	96.01	54.52	86.38	83.80
4	Italian (w/o Aug)	G+S	76.10	37.79	88.86	30.83	81.66	54.84
5	Adj Aug	G+S	80.86	42.48	90.02	33.19	84.56	58.95
6	Adv Aug	G+S	82.70	42.30	90.00	33.07	85.07	59.21
7	CN Aug	G+S	81.18	40.02	89.23	32.23	83.00	56.87
8	NE Aug	G+S	80.07	42.62	89.83	33.36	84.33	59.07
9	Verb Aug	G+S	80.15	39.99	89.48	31.90	83.10	57.04
10	Tense Aug	G+S	84.13	44.49	90.26	33.46	85.14	60.05
11	Comp Aug	G+S	85.98	45.12	90.56	34.54	85.66	61.66
12	IT5 [14], with Comp Aug	G+S	50.57	10.97	79.38	16.25	56.31	29.76
–	byT5 [24]	G+S+C	87.20	53.20	—	38.50	87.50	—
13	Italian (w/o Aug)	G+S+C	89.22	56.46	92.72	40.48	90.02	70.38
14	Adj Aug	G+S+C	89.46	56.77	92.90	40.49	90.02	70.66
15	Adv Aug	G+S+C	89.69	57.00	92.95	40.62	90.71	70.66
16	CN Aug	G+S+C	90.46	57.28	92.85	40.80	90.21	70.59
17	NE Aug	G+S+C	89.28	56.98	92.76	40.57	90.27	70.56
18	Verb Aug	G+S+C	**90.56**	56.15	92.80	40.49	90.10	70.46
19	Tense Aug	G+S+C	89.35	**57.48**	**92.97**	**40.95**	**90.97**	**70.88**
20	Comp Aug	G+S+C	89.44	56.58	92.79	40.87	90.21	70.63

strategy aids in understanding complex language patterns and semantic information; (iii) it demonstrates superior performance in spelling and pronunciation-sensitive tasks due to its resilience to noisy data; (iv) and as a token-free model, it operates directly on raw UTF-8 data. Importantly, byT5 has shown state-of-the-art results on multilingual NLP benchmarks [11, 12, 13]. We also conducted experiments with T5 specialized on ITalian (IT5) [14], a model that had demonstrated promising results in Italian language understanding and generation across various benchmarks.

Our fine-tuning strategy involves two stages: initial pre-fine-tuning with gold and silver (for exp.1–12), and gold, silver, and copper (for exp.13–20) data for 5 epochs to provide foundational DRS knowledge, followed by fine-tuning on only gold data—without augmentation—with an early stopping mechanism [15]. The hyperparameter setting used in our experimentation is listed in Table 5.

Evaluation Methods For evaluation, we employ distinct methods for semantic parsing and natural language generation tasks. In parsing evaluation, we first transform DRS into Penman notation [16], then use SMATCH [17] to calculate the overlap of triples between system output and the gold standard, assessing the output using F-Score to balance precision and recall [18]. For generation evaluation, we use a combination of different automatic metric evaluations including (i) n-gram-based measures like BLEU [19], METEOR [20], and chrF [21]; (ii) neural model-based COMET score [22]; and (iii) the pre-trained model-based BERT-Score ("bert-base-multilingual-cased" model) [23]. These comprehensive evaluations allow us to assess both the technical accuracy and the linguistic quality of our model output across parsing and generation tasks.

Results and Analysis The experimental results reported in Table 3 demonstrate the efficacy of diverse DA strategies in enhancing semantic parsing and text generation tasks for Italian DRS. We used different variants of T5 (byT5 and IT5) models and evaluated performance on the PMB-5.0.0 dataset, utilizing SMATCH F1 for parsing and BLEU, METEOR, COMET, chrF, and BERT-Score metrics for generation tasks.

In the multilingual baseline comparisons, Italian (76.10% SMATCH F1 for parsing) exhibits superior performance to Dutch (42.77%) and comparable results to German (73.00%), while expectedly trailing English (91.42%). For generation, Italian achieves baseline scores of 37.79 BLEU, 30.83 METEOR, 81.66 COMET, 54.84 chrF, and 88.86 BERT-Score, positioning it better than Dutch and German in all metrics.

Individual augmentation strategies uniformly yield improvements over the baseline Italian model. For parsing tasks, tense augmentation demonstrates the highest efficacy among singular strategies, achieving 84.13% SMATCH F1 (exp. 10). In generation tasks, tense augmentation emerges as the most effective individual strategy, attaining scores of 44.49 BLEU, 33.46 METEOR, 85.14 COMET, 60.05 chrF, and 90.26 BERT-Score (exp. 10). These enhancements indicate that each augmentation type contributes uniquely to the semantic understanding and generative capabilities of the neural model.

The effectiveness of tense augmentation correlates with the significant presence of temporal relations and structural simplicity in the test set's DRSs. Our analysis

reveals that approximately 94.05% of the test set contains active voice examples, while passive voice examples account for only 5.95%, making tense augmentation particularly valuable for improving model performance in sentence structures. Additionally, 98.20% of the test set consists of simple sentences, which further emphasizes the importance of augmentations that can enhance lexical diversity without overcomplicating sentence complexity. We observed the following distribution of sentence types in our test set: declarative (87.57%), exclamatory (2.52%), and interrogative (9.78%), reinforcing the need for augmentations that effectively handle these dominant structures.

The compound augmentation approach, which integrates all augmentation strategies, produces the optimal results for the Gold+Silver (G+S) dataset. This comprehensive strategy achieves 85.98% SMATCH F1 for parsing and notable improvements across all generation metrics (45.12 BLEU, 34.54 METEOR, 85.66 COMET, 61.66 chrF, and 90.56 BERT-Score), underscoring the synergistic benefits of combining diverse augmentation techniques (exp. 11). The performance of IT5 proved inadequate when applied to formal meaning representations i.e., DRS. The model exhibited suboptimal results in both semantic parsing and text generation tasks subsequent to fine-tuning on the compound augmentation dataset. The suboptimal performance of IT5 can be attributed to its pre-training focus on general Italian language tasks rather than formal meaning representations like DRS. This limitation highlights the challenges of adapting general-purpose language models to specialized semantic processing tasks.

Furthermore, comparisons with extant literature ([24] in Table 3) reveal the superior performance of our proposed approach. The referenced study reports 87.20% SMATCH F1 for parsing and 53.20 BLEU, 38.50 METEOR, and 87.50 COMET for generation on the Gold+Silver+Copper (G+S+C) dataset. In contrast, our Italian model (exp. 13—G+S+C baseline) achieves 89.22% SMATCH F1, 56.46 BLEU, 40.48 METEOR, 90.02 COMET, 70.38 chrF, and 92.72 BERT-Score on the same dataset, representing significant advancements across all metrics.

The most notable results are observed in the G+S+C dataset experiments. Verb Augmentation (exp. 18) achieves the highest parsing score of 90.56% SMATCH F1, while Tense Augmentation (exp. 19) leads in generation with scores of 57.48 BLEU, 40.95 METEOR, 90.97 COMET, 70.88 chrF, and 92.97 BERT-Score. These results not only surpass previous benchmarks but also approach the performance metrics of English, a high-resource language, despite comparatively limited lexical resources for Italian. The similar performance between the baseline Italian model (exp. 13) and compound augmentation (exp. 20) on G+S+C is primarily attributable to the substantial volume of Copper data (92, 394 examples). These Copper examples, which are Italian translations of the English Bronze dataset, outnumber our G+S compound augmentation by approximately 2:1, somewhat diminishing the observable impact of augmentation strategies. Furthermore, in our experiments with G+S+C (exp. 13–20), we have used the Copper version without any augmentation—just to have a fair comparison with literature reference (see experimental results of [24] in in Table 3). These experimental outcomes provide strong evidence that DA can significantly enhance the performance of semantic parsing and text generation models for Italian.

5. Conclusion

This study has successfully developed and evaluated a novel cross-lingual DA technique for Italian, specifically tailored for DRS-based semantic parsing and generation tasks. Our research has made significant improvements in addressing the challenges faced by low-resource languages in advanced NLP tasks. The proposed augmentation methodology, leveraging English WordNet to enhance Italian semantic datasets, has demonstrated remarkable effectiveness. Empirical evidence shows substantial improvements in performance scores for both DRS parsing and generation tasks in Italian. Notably, our approach achieved a 90.56% SMATCH F1 score for parsing and significant enhancements across all generation metrics (BLEU: 57.48, METEOR: 40.95, COMET: 90.97, chrF: 70.88, BERT-Score: 92.97) on the G+S+C dataset, surpassing both baseline models and previous state-of-the-art results. Our detailed analysis reveals that data augmentation positively affects the handling of Italian-specific linguistic features in semantic processing. The improvements observed across various augmentation strategies indicate enhanced capability in managing syntactic flexibility and grammatical nuances in Italian. This suggests a successful transfer of semantic knowledge through the lens of Italian DRS.

Limitations:
Despite our results approach the performance metrics of English—a rich resource language, there remains a gap that future research could address. For example, the original sentence "Tom è piuttosto scarso a tennis." ("*Tom is rather poor at tennis.*") becomes "Bob era piuttosto ricco con i single." ("*Bob was sort of rich at singles.*") While this method introduces linguistic diversity, it can result in less coherent sentences in some cases, as seen in this example. Such limitations are common with cross-lingual augmentation strategies through back-and-forth language translations, which focus on lexical variation over syntactic coherence. Future refinement, such as filtering improbable substitutions or adding human validation, could help ensure more consistent logicality in cross-lingual semantic tasks.

Acknowledgments

We thank "High-Performance Computing for Artificial Intelligence (HPC4AI) at the University of Turin" for providing GPU support [25].

References

[1] V. Basile, J. Bos, K. Evang, N. Venhuizen, Developing a large semantically annotated corpus, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 3196–3200. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/534_Paper.pdf.

[2] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, N. Schneider, Abstract meaning representation for sembanking, in: Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, 2013, pp. 178–186.

[3] M. S. Amin, L. Anselma, A. Mazzei, Exploring data augmentation in neural DRS-to-text generation, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 2164–2178. URL: https://aclanthology.org/2024.eacl-long.132.

[4] L. Abzianidze, R. van Noord, C. Wang, J. Bos, The parallel meaning bank: A framework for semantically annotating multiple languages, Applied mathematics and informatics 25 (2020) 45–60.

[5] M. S. Amin, A. Mazzei, L. Anselma, et al., Towards data augmentation for drs-to-text generation, in: CEUR WORKSHOP PROCEEDINGS, volume 3287, CEUR-WS, 2022, pp. 141–152.

[6] B. Li, Y. Wen, W. Qu, L. Bu, N. Xue, Annotating the little prince with chinese amrs, in: Proceedings of the 10th Linguistic Annotation Workshop held in Conjunction with ACL 2016 (LAW-X 2016), 2016, pp. 7–15.

[7] J. Bos, The sequence notation: Catching complex meanings in simple graphs, in: Proceedings of the 15th International Conference on Computational Semantics (IWCS 2023), Nancy, France, 2023, pp. 1–14.

[8] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of big data 6 (2019) 1–48.

[9] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for NLP, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 968–988. URL: https://aclanthology.org/2021.findings-acl.84. doi:10.18653/v1/2021.findings-acl.84.

[10] C. Wang, H. Lai, M. Nissim, J. Bos, Pre-trained language-meaning models for multilingual parsing and generation, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5586–5600. URL: https://aclanthology.org/2023.findings-acl.345. doi:10.18653/v1/2023.findings-acl.345.

[11] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, C. Raffel, Byt5: Towards a token-free future with pre-trained byte-to-byte models, Transactions of the Association for Computational Linguistics 10 (2022) 291–306.

[12] L. Stankevičius, M. Lukoševičius, J. Kapočiūtė-Dzikienė, M. Briedienė, T. Krilavičius, Correcting diacritics and typos with a byt5 transformer model, Applied Sciences 12 (2022) 2636.

[13] J. Belouadi, S. Eger, ByGPT5: End-to-end style-conditioned poetry generation with token-free language models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 7364–7381. URL: https://aclanthology.org/2023.acl-long.406. doi:10.18653/v1/2023.acl-long.406.

[14] G. Sarti, M. Nissim, IT5: Text-to-text pretraining for Italian language understanding and generation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9422–9433. URL: https://aclanthology.org/2024.lrec-main.823.

[15] R. van Noord, A. Toral, J. Bos, Character-level representations improve DRS-based semantic parsing even in the age of BERT, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4587–4603. URL: https://aclanthology.org/2020.emnlp-main.371. doi:10.18653/v1/2020.emnlp-main.371.

[16] R. T. Kasper, A flexible interface for linking applications to Penman's sentence generator, in: Speech and Natural Language: Proceedings of a Work-

shop Held at Philadelphia, Pennsylvania, February 21-23, 1989, 1989. URL: https://aclanthology.org/H89-1022.

[17] S. Cai, K. Knight, Smatch: an evaluation metric for semantic feature structures, in: H. Schuetze, P. Fung, M. Poesio (Eds.), Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 748–752. URL: https://aclanthology.org/P13-2131.

[18] W. Poelman, R. van Noord, J. Bos, Transparent semantic parsing with Universal Dependencies using graph transformations, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 4186–4192. URL: https://aclanthology.org/2022.coling-1.367.

[19] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[20] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[21] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049. doi:10.18653/v1/W15-3049.

[22] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213. doi:10.18653/v1/2020.emnlp-main.213.

[23] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[24] X. Zhang, C. Wang, R. van Noord, J. Bos, Gaining more insight into neural semantic parsing with challenging benchmarks, in: C. Bonial, J. Bonn, J. D. Hwang (Eds.), Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 162–175. URL: https://aclanthology.org/2024.dmr-1.17.

[25] M. Aldinucci, S. Rabellino, M. Pironti, F. Spiga, P. Viviani, M. Drocco, M. Guerzoni, G. Boella, M. Mellia, P. Margara, I. Drago, R. Marturano, G. Marchetto, E. Piccolo, S. Bagnasco, S. Lusso, S. Vallero, G. Attardi, A. Barchiesi, A. Colla, F. Galeazzi, Hpc4ai: an ai-on-demand federated platform endeavour, in: Proceedings of the 15th ACM International Conference on Computing Frontiers, CF '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 279–286. URL: https://doi.org/10.1145/3203217.3205340. doi:10.1145/3203217.3205340.

A. Data Transformation through Augmentation

The SBN is graphically shown in Figure 1 both with and without augmentation (a and b), highlighting the distinctions between proper nouns, common nouns, adjectives, adverbs, and verbal tense augmentations. With this augmentation, the original sentence "Tom è piuttosto scarso a tennis." or "Tom is rather poor at tennis." becomes "Bob era piuttosto ricco con i single." or "Bob was sort of rich at singles.". In Figure 1, augmented logical notions are highlighted conceptually. We used the Parallel Meaning Bank (PMB) dataset for this investigation, using both its gold (completely manually annotated) and silver (partially manually annotated) standard versions, and split it according to conventional methods for training, development, and testing.

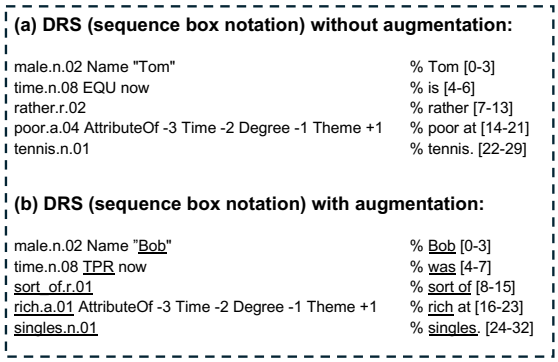

Figure 2: Graphical representations of DRS (a) without augmentation for the text "Tom è piuttosto scarso a tennis." or "Tom is rather poor at tennis." and (b) with augmentation for the text "Bob era piuttosto ricco con i single." or "Bob was sort of rich at singles.".

In order to provide transformed instances for neural semantic processing and text generation, named entities, lexical, and grammatical DA approaches were applied to the original sentences as shown in Table 4. It demonstrates how varying a sentence's constituent parts can improve dataset variety. When it comes to named entities, the sentence "Tom asked Mary if she had been to Boston" becomes "Bob asked Sarah if she had been to Cambridge", demonstrating how proper nouns are substituted. "Tom played with his dog" becomes "Tom played with his puppy" when it comes to common nouns, illustrating synonym replacement with hyponyms. Verb augmentation is demonstrated by changing the verb from "Tom thinks I stole the money" to "Tom philosophizes I stole the money", changing the meaning of the phrase. To demonstrate adjective and adverb augmentations, lexical entities are changed from "ill" to "well" and "deeply" to "profoundly", respectively. The last example of grammatical augmentation is when "A girl is playing the flute" is changed to one of three tenses: "A girl was playing the flute", "A girl will be playing the flute", or "A girl has been playing the flute". These illustrations show how enhancing various phrase constituents can produce diverse and richer datasets, supporting the creation of strong neural models.

B. Statistical distribution of examples

Table 1 reports the number of training, development, and testing examples in each language as well as the statistical distribution of the dataset used for multilingual baselines. Train Gold (T_Gold), Train Silver (T_Silver), Development (Dev), and Test sets comprise the dataset. There are 4,316 T_Silver, 555 Dev, 555 Test, and 745 T_Gold examples for Italian. There are 6,862 T_Silver, 900 Dev, 900 Test, and 1,206 T_Gold examples in German. There are 1,646 T_Silver, 435 Dev, 435 Test, and 586 T_Gold examples in Dutch. There are 143,731 T_Silver, 1,132 Dev, 1,132 Test, and 9,057 T_Gold examples for English, the language with the largest representation. As can be seen from this distribution, the English corpus is substantially larger than the other languages, offering a solid dataset for training and evaluation. This diversity in dataset size across languages highlights the varying amounts of linguistic data available for training multilingual models.

C. Impact of Augmentation on Dataset Size

Table 2 compares the number of instances with and without augmentation to those with individual and compound augmentations to show how different augmentation methods affect the size of the dataset. Without any augmentation, the original dataset had 5061 gold-silver samples altogether, 4316 silver examples, and 745 gold examples. Applying individual augmentations, including Named Entities, Common Noun, Adjective, Adverb, and Verb augmentations, twice the size of the dataset; for every augmentation type, there are 1490 gold, 8632 silver, and 10122 gold-silver examples. Even more so, tense augmentation quadruples the amount of the dataset to 2980 gold, 17264 silver, and 20244 gold-silver examples. Compound augmentation yields the largest gain, ninefolding the dataset size to 6705 gold, 38844 silver, and 45549 gold-silver examples. Compound augmentation incorporates numerous augmentation strategies. The number of examples in both the development and test sets stays at 555. This notable augmentation of the dataset size highlights the potential for more comprehensive and diverse

Table 4
Named-entities, lexical, and grammatical DA approaches for neural semantic parsing and text generation. The English translation is mentioned in double quotes.

Augmentation Type	Original Examples	Transformed Examples
Named Entities	Tom ha chiesto a Mary se fosse stata a Boston. "Tom asked Mary if she had been to Boston."	Bob ha chiesto a Sarah se fosse stata a Cambridge. "Bob asked Sarah if she had been to Cambridge."
Common Noun	Tom ha giocato con il suo cane. "Tom played with his dog."	Tom ha giocato con il suo cucciolo. "Tom played with his puppy."
Verb	Tom pensa che io abbia rubato i soldi. "Tom thinks I stole the money."	Tom filosofeggia che ho rubato i soldi. "Tom philosophizes I stole the money."
Adjective	Lui è malato. "He is ill."	Lui è bene. "He is well."
Adverb	La ragazza è profondamente legata a sua zia. "The girl is deeply attached to her aunt."	La ragazza è sinceramente legata a sua zia. "The girl is sincerely attached to her aunt."
Grammatical	Una ragazza suona il flauto. "A girl is playing the flute."	Una ragazza suonava il flauto. "A girl was playing the flute." Una ragazza suonerà il flauto. "A girl will be playing the flute." Una ragazza ha suonato il flauto. "A girl has been playing the flute."

training data, which can enhance the robustness and performance of neural networks.

D. Hyperparameters For Experimental Implementation

In Table 5, we report a list of the main hyperparameters used in our experimental implementation. We have used the same experimental setting for all of our experiments reported in Table 3. We used the AdamW optimizer with a batch size of 32, a learning rate of 1e-4, and a maximum sequence length of 512 tokens. Throughout our experiments, we used GeGLU for activation functions. Two rounds of fine-tuning were carried out: the first stage lasted for five epochs, and the second stage used early stopping criteria to dynamically decide the ideal number of epochs depending on metrics related to the performance of the model. These hyperparameters were chosen with attention to guarantee reliable operation and efficient byT5 model customization to our particular tasks and datasets.

Table 5
Hyperparameter setting for our experiments.

Parameter	Value
Optimizer	AdamW
Learning rate	1e-4
Batch size	32
Max length	512
Activation function	GeGLU
Epoch for fine-tuning stage 1	5
Epoch for fine-tuning stage 2	early stopping