Exploring the Dissociated Nucleus Phenomenon in Semantic Role Labeling

Anonymous ACL submission

Abstract

Dependency-based Semantic Role Labeling (SRL) is bound to Dependency Parsing, as the arguments of a predicate are identified through the token that heads the dependency relation subtree of the argument span. However, most dependency-based SRL corpora are susceptible to the *dissociated nucleus* problem: when a subclause's semantic and structural cores are two separate words, the dependency tree chooses the structural token as the head of the subtree, coercing the SRL annotation to make the same choice. This leads to undesirable consequences: when directly using the output of a dependency-based SRL method in downstream tasks it is useful to work with the token representing the semantic core of a subclause, not the structural core. In this paper, we carry out a linguistically-driven investigation on the *dissociated nucleus* problem in dependency-based SRL and propose a novel algorithm which aligns predicate-argument structures to the syntactic structures from Universal Dependencies to select the semantic core of an argument. Our analysis shows that *dissociated nuclei* appear more often than one could expect, and that our novel algorithm greatly enhances the richness of the semantic information in dependency-based SRL. We release the software to reproduce our experiments at `http://omitted.link`.

1 Introduction

Within the field of Natural Language Processing, Semantic Role Labeling (Gildea and Jurafsky, 2002, SRL) aims to recognize the semantic information conveyed by a sentence, more specifically identifying *who* did *what* to *whom*, *where* and *how* (Màrquez et al., 2008). Over the years, SRL has split into two main annotation formalisms, namely span-based and dependency-based. The key difference between the two is how they identify the roles of a predicate: span-based SRL directly identifies a span of the input text as the argument of a predicate, whilst in dependency-based SRL the word that heads the syntactic dependency relation subtree corresponding to the argument span is the one identified as the argument. Using dependency-based SRL can be beneficial in real-world settings, as i) dependency-based SRL parsers have achieved greater results on standard benchmarks, and ii) the head words of the SRL arguments can be directly utilized in downstream tasks, such as Entity Linking, Word Sense Disambiguation (WSD), Coreference Resolution, and Knowledge Base construction. However, this step requires the foregoing head words to carry the "semantic meaning" of the role.

This is often not the case when examining both the output of state-of-the-art dependency-based SRL systems and the corpora they were trained on, such as CoNLL-2009 (Hajič et al., 2009). In these annotations, it is not uncommon to have an adpositional clause serving as the head word of a semantic role, even though adpositions do not represent the 'semantic core' of that role. In linguistics, this phenomenon is referred to as the *dissociated nucleus* (Tesnière, 2015, ch. 23). Although it encompasses many different syntactical constructions, we focus on adpositional clauses present in CoNLL-2009 across all of its languages.

In this short paper, we carry out a linguistically-driven investigation on dissociated nuclei in dependency-based SRL, uncovering the extent of this problem and how it affects the semantic aspect of this task in English and other languages. Moreover, we introduce SemDepAlign, a simple yet effective algorithm, which is capable of mitigating this phenomenon significantly by aligning predicate-argument structures in SRL with syntactic parses from the Universal Dependencies project, increasing the quantity of semantic cores in CoNLL-2009. Finally, we provide a glimpse at how addressing dissociated nuclei simplifies the alignment between SRL parses and AMR graphs. We release SemDepAlign and Aligned-CoNLL09

– the result of applying SemDepAlign to CoNLL-2009 – in the hope that our work can encourage "more semantic" SRL and foster future integration of this task into downstream applications.

2 SRL and Dependency Parsing

The interconnection between the tasks of SRL and Dependency Parsing is clear: both investigate how words in the same sentence relate to each other, respectively in a semantic and syntactic sense. The Conference on Computational Natural Language Learning (CoNLL) has organized several Shared Tasks regarding both tasks, culminating with the CoNLL-2008 Shared Task (Surdeanu et al., 2008) that asked participants to identify both types of relations within an English-only corpus. This task can be seen as the first occurrence of dependency-based SRL, as it explicitly tied the SRL annotations to the dependency relation tree of the sentence.

Considering that CoNLL-2008 and its multilingual successor CoNLL-2009 (Hajič et al., 2009) are among the most utilized training sets for dependency-based SRL systems, the importance of the dependency annotation scheme that they use has been somewhat neglected across the literature, as it influences the output of these systems. If one inspects the output of a system trained on CoNLL-2009, they can expect to encounter the dissociated nucleus phenomenon (Tesnière, 2015[1959], ch. 23), a linguistic phenomenon present in the training dataset. For example, the English sentence "That is a service to the nation" presents a dissociated nucleus, i.e. the structural and semantic functions of the subclause "to the nation" are fulfilled by two separate tokens, 'to' and 'nation' respectively. The annotation provided within CoNLL-2009 identifies the syntactic core 'to' with the argument `A2` for the predicate 'service' because it is the head of the original dependency relation subtree corresponding to the argument span. Consequently, many tokens annotated as arguments are simple adpositions of little semantic significance, e.g. if we wanted to extract relations or perform disambiguation on the example above, we would have much more interest in focusing on the word 'nation' than the adposition 'to'. Therefore, we argue that both the training corpora and dependency-based SRL systems should identify the semantic core of an argument span as the head of the argument. We provide further examples of this phenomenon in non-English partitions of CoNLL-2009 in Appendix A.

3 Re-associating Dissociated Nuclei

Having established that the current annotations in CoNLL-2009 are susceptible to the dissociated nucleus phenomenon, we aim to mitigate this issue by introducing a subtree alignment algorithm that leverages the characteristics of Universal Dependencies (de Marneffe et al., 2021, UD) to collapse arguments that have been placed on structural tokens with their corresponding semantic tokens. UD explicitly addresses the dissociated nucleus issue by extending the definition of a nominal to encompass the entire nominal extended projection, following the linguistic theory proposed by Grimshaw (1990). The nominal head is used as the referential core and the adposition is treated as a functional marker (de Marneffe et al., 2021, Section 3.1.1). When constructing the dependency tree structures, UD guidelines (de Marneffe et al., 2021, Section 2.1.1) indicate that the head of a particular subclause should be its content word, i.e. the nominal head. Parsers trained on UD Treebanks recognize dependency subtrees where the head is the semantic core of the subclause, effectively mitigating the dissociated nucleus phenomenon. We leverage this characteristic of UD parsers to automatically annotate the whole CoNLL-2009 corpus using `trankit` (Nguyen et al., 2021), which emerges as the strongest UD parser in the comparison we include in Appendix B.

3.1 SemDepAlign: subtree alignment

We introduce SemDepAlign, a novel algorithm for syntactic parse semi-alignment from the dependency annotations in CoNLL-2009 to UD, described in Algorithm 1. SemDepAlign is a deterministic subtree aligning algorithm that, for each role token t in a semantic frame, attempts to find the UD subtree that most closely matches the original subtree headed by t in the original dependency tree of CoNLL-2009. It then returns the head node t' of the UD subtree, which will be assigned the role label in the aligned SRL annotation.

As shown in Algorithm 1, SemDepAlign starts from the UD root node (*root_ud*), loops over the nodes of the tree through a breadth-first search (BFS), and finds the node which heads the subtree with minimal symmetric set difference (*SymDiff*) with the set of tokens in the original role span (*role_tokens*). Intuitively, if the symmetric difference between the original and the UD subtree is the empty set, they share all elements in the span

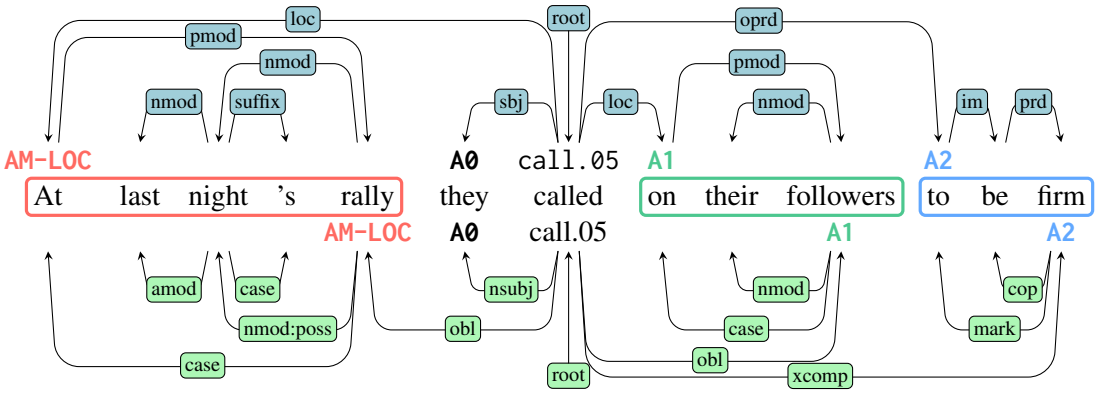

Figure 1: An edited example from the English CoNLL-2009 development set: original (top) and aligned (bottom) dependency and role annotations for the predicate *called*. We represent role annotations through colored clusters, where SemDepAlign aligns the head token to a more semantic token heading the UD subtree closest to the original.

Algorithm 1: SemDepAlign

input: the role node *role_node*; the root node of the UD dep-tree *root_ud*.
output: the head node of the role in the UD dep-tree.

role_tokens ← *get_tokens*(*role_node*)
ud_role_subtree ← *root_ud*
min_nodes ← *SymDiff*(*get_tokens*(*root_ud*), *role_tokens*)
for *node* ← *BFS*(*root_ud*)**:**
  subtree_tokens ← *get_tokens*(node)
  extra_nodes ← *SymDiff*(*subtree_tokens*, *role_tokens*)
  min_nodes ← min(*min_nodes*, *extra_nodes*)
return *ud_role_subtree*

and we can extract the head of a subtree of the UD dependency tree that matches exactly with the role span in the original dependency tree. Otherwise, selecting the head of the UD subtree with minimal symmetric difference with the original subtree is equivalent to selecting the subtree with the most overlap with the original span.

Figure 1 gives an example of the output of SemDepAlign: the top part of the figure shows the original annotation of the sentence in English CoNLL-2009, with the presence of a dissociated nucleus in three of the four roles for the predicate "called" (highlighted in color), and the bottom part shows the output of our alignment procedure, which moves the role annotations to the tokens that perform the semantic function.

3.2 Aligned-CoNLL09: analysis

We apply SemDepAlign to CoNLL-2009 to mitigate the dissociated nucleus phenomenon, obtaining the Aligned-CoNLL09 dataset. The number of role tokens that were aligned is considerable over all CoNLL-2009 languages (between 21% and 32% of the total roles), except for Czech (~7%). We consider the original tokens that were modified by SemDepAlign and the resulting aligned role tokens. We measure three metrics on these two sets: the number of **content words**, i.e. tokens that are either nouns, adjectives, adverbs, or verbs; the number of **unique tokens** that appear as role tokens; and the number of **unique synsets** that are disambiguated from role tokens, according to a WSD system (Orlando et al., 2021). In Table 1 we show how SemDepAlign dramatically increases the semantic content of role tokens in English, Spanish and German, identifying more than $4\times$ times the amount of content words, more than $2.5\times$ times the number of unique tokens and around $3\times$ times the unique synsets. We find a smaller but consistent increase of semantic content in Catalan and Chinese, whilst in Czech all metrics are similar, indicating a reduced effect of SemDepAlign.

4 Integrating re-associated nuclei

Although we demonstrate that re-associated nuclei in dependency-based SRL provide additional semantic information, an important research question is whether integrating our proposal into current systems can lead to a decrease in performance. Therefore, we build on top of the strong model proposed by Conia and Navigli (2020) and design a new model that jointly learns both types of role annotations, i.e. the original role tokens and the aligned ones. In brief, the original architecture derives a predicate- and argument-specific representation for each word in a sentence through a pretrained BERT-like language model (Devlin et al., 2019, PLM) and several "fully-connected"

Language	CA		CS		DE		EN		ES		ZH	
Dataset	O	A	O	A	O	A	O	A	O	A	O	A
# modified roles	3356 (29.1%)		3578 (7.3%)		314 (26.9%)		4205 (30.3%)		3756 (32.4%)		4021 (21.7%)	
Content words	1159	1344	2486	2923	59	246	680	2713	574	3019	595	618
Unique tokens	529	1952	2574	2473	108	274	477	1906	563	2324	1223	1742
Unique synsets	825	921	1339	1397	54	191	481	1387	457	1708	219	266

Table 1: Semantic variety of role tokens that were modified when aligning the original CoNLL-2009 (O) to Aligned-CoNLL09 (A).

System	Validation			Test		
	P	R	F1	P	R	F1
Baseline	91.18	91.55	91.37	91.95	92.38	92.16
Aligned	91.38	91.33	91.35	92.07	92.25	92.16

Table 2: Results on the validation and test sets of CoNLL-2009 (English).

System	Predicate F1	Role F1	Aligned Role F1
Baseline	95.15	89.75	-
Aligned	95.22	89.85	87.85

Table 3: Finer-grained evaluation on the English CoNLL-2009 test set on predicates, roles and aligned roles.

stacked-BiLSTM sequence encoders. It then derives a score distribution over the role vocabulary through a simple linear classifier applied over a further predicate-specific encoding of each word. The model is trained to minimize the sum of categorical cross-entropy losses on predicate identification, predicate disambiguation and argument labeling.

To adapt this model for our joint modeling task, we duplicate the linear classifier for the semantic roles and set two different targets for the two role classifiers: the original role token and label from CoNLL-2009 and the aligned role token and label obtained with SemDepAlign. Our final loss adds terms for UD-aligned argument identification and classification to the original loss.

Experimental setup We use XLM-RoBERTa-base (Conneau et al., 2020) as the underlying PLM, and leave hyperparameters unrelated to our experimental setup unchanged. We conduct our experiments on the English split of CoNLL-2009.

Results Table 2 compares the results of our joint-modeling alignment system against our baseline on the English CoNLL-2009 validation and test sets. Importantly, we observe that the additional task of modeling the semantic core of an argument does not significantly alter the performance (equal F1 score on the test), despite the added difficulty brought by the identification of semantic cores. Table 3, instead, provides a breakdown of the F1 scores on predicate, role and aligned role predictions. The aligned system slightly improves over the baseline despite being tasked with a more complex objective, although it can exploit extra annotated data (the aligned roles). More interestingly, we observe that the F1 score on the semantic heads is comparable, indicating that the model is able to identify UD-aligned roles effectively. When extending this experimental setup to all languages of CoNLL-2009 (Appendix C) we reach very similar results and conclusions.

5 Related work

Syntactic information has always been considered important to recognizing semantic frames in SRL. Marcheggiani and Titov (2017) were among the first to model the dependency information provided in dependency-based SRL, followed most recently by Xia et al. (2020); Fei et al. (2021). These works differ in modeling choices and in what kind of extra syntactic data to include (including constituency trees, and POS tags). We considered other syntactic frameworks, such as HPSG (Pollard and Sag, 1994), to align the role annotations. HPSG robustly models the relationship between semantic cores of a sentence, but the lack of automatic tools with an acceptable performance and the difficulty in aligning dependency-based subtrees to HPSG spans compelled us to use UD.

6 Conclusion

In this paper, we conducted an in-depth investigation on the dissociated nucleus issue in dependency-based SRL. We introduced SemDepAlign, a novel alignment method to leverage syntactic information to re-align semantic roles in dependency-based SRL formalisms. Our analyses and experiments in SRL modeling prove that our approach to dissociated nuclei brings more semantic richness whilst remaining competitive on standard benchmarks.

7 Limitations

A limitation of our work is that it builds upon existing dependency parsers trained on Universal Dependencies. These parsers have reached high robustness across many languages, between 85 and 93 in Labeled Attachment Score (LAS) on the languages present in CoNLL-2009. But the error that these automatic methods necessarily encounter propagates directly to our alignment algorithm, with no way to recover from the mistake. This limitation would be even more impactful in languages where the automatic dependency parser performed worse, presumably in low-resource settings, preventing a robust expansion of our work to these settings.

A more methodological limitation of our contributions concerns the availability of the CoNLL-2009 dataset. Although it is a corpus well-established in the SRL literature, it has a proprietary licensing scheme and one must acquire the resource from the Linguistic Data Consortium (LDC). We trust that, given the importance of the corpus, this will not limit the relevance of our work.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Liang Chen, Peiyi Wang, Runxin Xu, Tianyu Liu, Zhifang Sui, and Baobao Chang. 2022. ATP: AMRize then parse! enhancing AMR parsing with PseudoAMRs. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2482–2496, Seattle, United States. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. 2021. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, Online. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

Jane Grimshaw. 1990. *Argument Structure*. The MIT Press, Cambridge, MA.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18. Association for Computational Linguistics.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark. Association for Computational Linguistics.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic Role Labeling: An Introduction to the Special Issue. *Computational Linguistics*, 34(2):145–159.

Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

C. Pollard and I.A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. University of Chicago Press.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177, Manchester, England. Coling 2008 Organizing Committee.

Lucien Tesnière. 2015. *Elements of Structural Syntax*. John Benjamins.

Qingrong Xia, Rui Wang, Zhenghua Li, Yue Zhang, and Min Zhang. 2020. Semantic role labeling with heterogeneous syntactic knowledge. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2979–2990, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Dissociated nuclei in non-English samples of CoNLL-2009

A.1 Catalan

Original sentence:
"Piqué recomana les fusions entre empreses per millorar la rendibilitat."
Translation:
"Piqué recommends mergers between companies to improve profitability."
Dissociated nucleus:
In the clause "per millorar" ("to improve"), 'per' ('to') is tagged as `argM-fin` for predicate 'recomana' ('recommends') instead of the head of the subclause 'millorar' ('improve').

A.2 German

Original sentence:
"Setzt Hessen auf eine Effizienzsteigerung der Verwaltung durch neue Steuerungsinstrumente ."
Translation:
"Hesse is focusing on increasing the efficiency of administration through new control instruments."
Dissociated nucleus:
Considering the predicate 'setzt' ('focus'), the clause "auf eine Effizienzsteigerung" ("on increasing the efficiency") is annotated with role `A1` on the token 'auf' instead of the semantic core 'Effizienzsteigerung'.

A.3 Spanish

Original sentence:
"Don Antonio se encontraba en su casa cuando sonó el timbre de la puerta."
Translation:
"Don Antonio was at his home when the doorbell rang."
Dissociated nucleus:
The role "en su casa" ("at his home") for predicate 'encontraba' ('was') is tagged as `arg2-loc` on the token 'en' ('in') instead of the semantic nucleus 'casa' ('home').

A.4 Chinese

Original sentence:
巴拉克 在 民意 测验 中 一直 表现 不 佳 。
Transliteration:
"Barak in public opinion test in continuously performance no good."
Translation:
"Barak has consistently underperformed in the polls."
Dissociated nucleus:
In the clause 在 民意 测验 ("in the public opinion polls") for the nominal predicate 佳 ('good'), as the token 在 ('in') is tagged as the `LOC` role, instead of the more semantic 测验 ('polls').

B Universal Dependency parsers

We consider three among the best off-the-shelf dependency parsers, namely `trankit` (Nguyen et al., 2021), `UDPipe` (Straka, 2018) and `Stanza` (Qi et al., 2020). Table 4 compares the reported evaluation of each parser on standard treebanks for Catalan, Czech, German, English, Spanish and Chinese. We choose `trankit` as it achieves a higher UAS and

Treebank	System	UAS	LAS
Catalan AnCora	trankit	**95.15**	**93.83**
	UDPipe	94.92	93.43
	Stanza	93.55	91.66
Czech PDT	trankit	**95.24**	**93.65**
	UDPipe	95.01	93.64
	Stanza	92.22	90.18
German GSD	trankit	**89.01**	**85.20**
	UDPipe	87.04	83.20
	Stanza	85.80	81.80
English EWT	trankit	**91.29**	**89.4**
	UDPipe	90.71	88.81
	Stanza	88.90	86.77
Spanish AnCora	trankit	93.29	91.10
	UDPipe	**93.68**	**91.92**
	Stanza	93.09	91.30
Chinese Simplified GSD	trankit	**87.38**	**84.82**
	UDPipe	72.74	70.28
	Stanza	73.41	70.65
Average	trankit	**91.89**	**89.67**
	UDPipe	89.02	86.88
	Stanza	87.83	85.40

Table 4: Performance of multiple off-the-shelf dependency relation parsers, measured by the standard Unlabeled and Labeled Attachment Scores (UAS and LAS). Boldface scores indicate the best performing system on a specific treebank.

LAS than the two alternatives in all languages except Spanish (slightly worse than UDPipe), with a considerable margin in Chinese.

C Multilingual experiments

We report the results of repeating the exact experimental setup of Section 4 on all languages included in CoNLL-2009. Table 5 showcases the metrics obtained through the official scoring script for our baseline and aligned method. In general, the aligned method performs very slightly worse than our baseline, losing at most 1 F1 point on the Catalan test set.

Table 6 reports the finer-grained evaluation carried out manually for each CoNLL-2009 test set over all languages. Again, we notice very similar Role F1 across all languages, indicating that tasking the model to recognize semantic roles does not have a strong effect on the normal role classification. It is also interesting to note that the Aligned role F1 score oscillates between 2-3.65 points under the Role F1 metric, so our aligned model can recognize semantic roles within the same performance ballpark.

Lang		Validation			Test		
		P	R	F1	P	R	F1
ca	B	87.97	87.76	87.86	88.12	88.04	88.08
	A	87.46	87.01	87.23	87.16	86.88	87.02
cs	B	86.49	86.38	86.44	86.18	86.14	86.16
	A	86.42	86.44	86.43	86.19	86.20	86.20
de	B	90.52	90.72	90.62	89.82	90.26	90.04
	A	90.74	90.32	90.53	89.63	90.02	89.82
en	B	91.18	91.55	91.37	91.95	92.38	92.16
	A	91.38	91.33	91.35	92.07	92.25	92.16
es	B	86.79	86.92	86.85	86.20	85.65	85.93
	A	86.70	86.41	86.55	85.89	85.18	85.53
zh	B	89.46	88.97	89.22	89.47	88.81	89.14
	A	89.24	89.06	89.15	89.28	88.66	88.97

Table 5: Results on the validation and test sets of all languages in CoNLL-2009. 'B' indicates the baseline models' results, whilst 'A' indicates the results achieved by our aligned version.

Lang		Predicate F1	Role F1	Aligned Role F1
ca	B	98.79	87.45	-
	A	98.67	86.64	82.99
cs	B	99.38	89.55	-
	A	99.39	89.59	87.11
de	B	94.88	89.42	-
	A	94.55	89.64	86.42
en	B	95.15	89.75	-
	A	95.22	89.85	87.85
es	B	99.00	86.33	-
	A	98.99	85.57	81.93
zh	B	96.17	86.06	-
	A	96.05	85.88	83.15

Table 6: Finer-grained evaluation on all CoNLL-2009 test sets on predicates, roles and aligned roles. 'B' indicates the baseline models' results, whilst 'A' indicates the results achieved by our aligned version.

D Evaluation through AMR

We develop an evaluation method based on the Abstract Meaning Representation formalism (Banarescu et al., 2013, AMR). The interconnection between SRL and AMR is well-known across the literature (Chen et al., 2022): both tasks aim to construct a semantic representation of a sentence, although SRL, covering only surface-level semantic frames, is more superficial than AMR, which aims to provide a more complete and in-depth structured representation that can interconnect different semantic frames.

We devise the **AMR-precision** metric: given a sentence S and its golden annotated AMR graph G_{AMR} with token-node alignments available and a set of dependency-based SRL predictions, we compute the ratio between the number of role tokens that are connected to their predicate in the AMR

Test dataset	Standard	Aligned	Δ
LORELEI	65.98	71.33	5.35
Weblog and WSJ	64.07	70.13	6.06
Xinhua MT	67.92	75.68	7.76
BOLT DF MT	60.92	68.17	7.25
BOLT DF English	56.22	62.12	5.90
Proxy reports	24.50	22.40	-2.10

Table 7: AMR-precision metric over standard and aligned role predictions derived from test datasets in AMR3.0. Δ indicates the difference in precision between the unified roles and the standard ones

graph over the total number of roles predicted. We only consider the predicted semantic frames where the predicate is present in the AMR graph.

Given the SRL system introduced in Section 4, we apply it to the AMR3.0 (LDC2020T02[1]) test datasets, keeping both the standard and the aligned role predictions. We then compute the AMR-precision for both sets of predicted roles, and compare them in Table 7. It is clear that aligned roles are more likely to be present in the corresponding AMR graph of a sentence, with a consistent difference in AMR-precision in all test datasets except *Proxy reports*. This particular dataset has a "templatic, report-like structure" as mentioned in the AMR3.0 guidelines, so it is possible that the reduced performance is due to this particular characteristic.

[1] catalog.ldc.upenn.edu/LDC2020T02