Improving Online Job Advertisement Analysis via Compositional Entity Extraction

Anonymous ACL submission

Abstract

We propose a compositional entity modeling framework for requirement extraction from Online Job Advertisements (OJAs), reframing the task from token classification to joint entity and relation classification to capture complex, multi-component requirement structures. Using an annotated dataset of 500 German OJAs, our empirical analysis reveals the prevalence of conjoined requirement structures and the importance of modeling complex semantic relationships between requirement components. Transformer-based models trained on our data achieve F1-scores of 0.856 for entity extraction and 0.911 for relation classification, demonstrating the effectiveness of our approach. This framework offers analytical benefits for labor market research and applications like skills monitoring or job-to-candidate matching, and we release our dataset to foster further research.

1 Introduction

Online Job Advertisements (OJAs) serve as a critical data source for understanding labor market dynamics across disciplines such as labor market research, education, and human resources (Khaouja et al., 2021). They offer detailed and up-to-date insights into in-demand skills, required qualifications, and evolving industry trends. By analyzing OJAs, researchers can identify skill gaps, changes in skill requirements and contribute to optimizing educational programs (Lima et al., 2018; Giabelli et al., 2021; Buchmann et al., 2022; Atalay et al., 2020, 2023). Job Ads have also been used in recruiting research (Castilla and Rho, 2023; Kim and Angnakoon, 2016) and for developing job recommendation systems via cv matching (Ntioudis et al., 2022; Smith et al., 2021; Belloum et al., 2019).

Work on Information Extraction (IE) in OJAs has mostly focused on skills extraction (see survey by Senger et al., 2024). Work extracting other information includes job tasks (Atalay et al., 2018, 2020, 2023), job titles (Baskaran and Müller, 2023; Li et al.; Giabelli et al., 2021; Rahhal et al., 2023), work tools (Güntürk-Kuhl et al.) and formal qualifications (Brown and Souto-Otero, 2020; Müller; Schimke, 2023; Börner et al., 2018). Collectively, these entities can be summarized as *requirements*, reflecting aspects of the position sought that pertain to the candidate.

Limitations of span-based approaches. A fundamental limitation of current approaches is the tendency to treat job-related requirements as standalone units rather than as interdependent components of a broader requirement framework. However, requirements in OJAs are often expressed as complex phrases with multiple components that interact with each other.

Figure 1 (upper part) illustrates this limitation using the example text "designing and implementing scalable systems using Java". As the Figure shows, there are multiple ways to annotate the requirements in this sentence. Traditional methods might extract "designing" or "implementing scalable systems" as isolated skills while failing to link them to the associated work tools ("Java"). This results in a loss of critical semantic dependencies.

Furthermore, "designing" in this context is not an independent skill but is intrinsically tied to "scalable systems". The requirement does not refer to design in a general sense, such as logo design or UI/UX design, but specifically to the conceptualization and architectural structuring of scalable systems. Nguyen et al. (2024) show that in four out of six skills datasets, between 16% and 22% of extracted spans are conjoined expressions, highlighting the prevalence of such structures in job advertisements.

Additionally, the extended example in the appendix (Figure 5) shows the full structure of the sentence, adding an experience level "initial experience" and the alternative between "Java" and "Python". A single-span extraction approach might

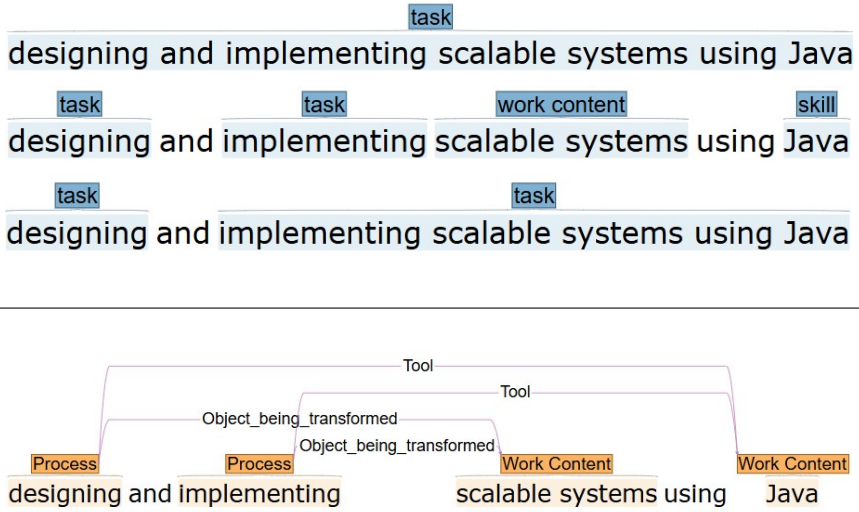


Figure 1: Comparison of traditional labeling schemes (top) and our proposed compositional approach (bottom).

fail to capture such logical relationships and thus at least miss semantic differentiation. At worst the meaning of entities could be misinterpreted.

Ultimately, traditional extraction approaches must navigate a trade-off: longer spans may capture more context but become more ambiguous and difficult to extract (Zhang et al., 2022a), while shorter spans risk omitting crucial relational details.

Contributions. To address these challenges, we propose a compositional entity modeling framework that decomposes requirement descriptions into their constituent components and explicitly models their relationships. Consequently, we methodologically reframe the task of requirements extraction from a token classification task to a combined token and relation classification task.

Figure 1 presents a side-by-side comparison that illustrates our approach: on the top, conventional entity extraction approaches fail to capture key dependencies, whereas on the bottom, the compositional modeling framework successfully extracts and links all relevant components. In more detail, our contributions are:

- Our framework for requirement extraction in OJAs addresses the limitations of conventional single-span entity extraction by explicitly modeling requirement components and their relationships.
- We evaluate the feasibility and effectiveness of our approach through (i) descriptive data analysis highlighting structural patterns in job requirements, (ii) conducting an exemplary demonstration revealing how the detailed structure of our framework adds analytical benefits, and (iii) benchmark evaluations of transformer-based models trained on our dataset.
- We present GOJA[1], a manually annotated gold standard dataset of 500 German job advertisements for our framework.

Our findings demonstrate that compositional modeling captures requirement dependencies more effectively than traditional approaches.

2 Our Approach

In this section, we present our overall procedure for establishing and applying a novel annotation schema.

2.1 Proposed Annotation Schema

The key observation underlying our approach is that fuzzy concepts such as skills and tasks, i.e. central job requirements, are often not directly represented in text as discrete, self-contained entities. Instead, they emerge compositionally from smaller, interrelated components. Our framework formalizes this by analyzing skills and tasks as chains of atomic entities linked by relations.

Table 1 provides a full overview of all 8 entity and 11 relation types in our annotation framework. In the following, we illustrate how the combination of atomic entities and relations forms job requirements in OJAs.

[1]We release the created GOJA to the research community upon acceptance of the paper.

Entity Type	Description	Example
Attitude	Indicates traits or dispositions desired in candidates.	You are adaptable
Attribute	Provides additional specifications about other entities.	You design logos for our customer
Experience Level	Indicates the level of knowledge or skills required.	Experience in Python
Formal Qualification	Identifies certifications or official qualifications required.	Bachelor's degree in Economics
Industry	Defines the industry or sector associated with the job.	You bring relevant experience in the automotive industry
Occupation	Specifies the role or position advertised.	We looking for a baker (m/f/d)
Process	Represents actions or sequences required to perform tasks.	You design Logos
Work Content	Describes the object or tool related to a task.	You design logos
Relation Type	**Description**	**Example (the relation connects the underlined entities)**
Alternative	Denotes alternatives between entities.	Bachelor's degree or minimum of three years professional experience)
Coordination	Connects coordinated morphems within sentences.	You pre- and post- process texts.
Degree of Autonomy	Specifies the level of autonomy in task execution.	You help your supervisor prepare presentations
Detail	Illustrates subcategories or specifics of an entity.	You are experienced with at least one programming language like Python
Negation	Highlights excluded processes or tasks.	This role does not include care duties.
Object Being Transformed (OBT)	Links processes to the items or entities they affect.	You design new logos
Related Entity Parts (REP)	Links separated parts of an entity.	You set the annual budget up
Specialization	Adds specificity to qualifications or roles.	A Bachelor's degree in Economics
Tool	Connects processes to the tools or methods used.	You design logos using Illustrator
Urgency	Indicates the importance or necessity of an entity.	Experience in Python is mandatory
Zero Relation	Used where the relation is self-evident.	You bring experience in programming

Table 1: Overview of Entity and Relation Types.

Tasks. Tasks are discrete units of work, defined as activities that transform inputs into outputs within an economic context (Autor and Handel, 2013; Rodrigues et al., 2021). As demand-side features of a job, tasks can be arbitrarily detailed or holistic. The PROCESS entitiy represents the action being performed, while the WORK CONTENT entity provides the context or target of the action. These components of tasks are linked through relations, which encode the semantic dependency between the process and the work content. Work contents can take the form of OBJECTS BEING TRANSFORMED (OBTs) i.e. things, immaterial objects, living objects or *work tools* (Fana et al., 2023), depending on whether the item is the subject of the process or the means by which the process is performed.

We refer to Figure 2 for an illustration: rather than directly annotating "designing scalable systems" as a task, we annotate the components "designing" as an atomic entity of type PROCESS and "scalable systems" as an entity of type WORK CONTENT. The OBT relation connects these two entities to form a task.

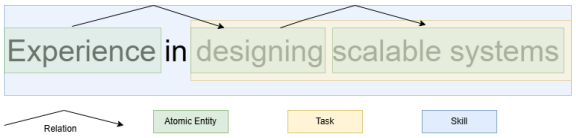


Figure 2: Example of analysis chains for skills and tasks.

Skills. Skills, on the other hand, are defined as the ability to perform a task effectively (Rodrigues et al., 2021). As skills pertain to job candidates, they represent the supply-side. Within our frame-

work, skills can be modeled as tasks augmented by entities of type EXPERIENCE LEVEL. Figure 2 shows how the task "designing scalable systems" plus the entity "Experience" form a skill. This skill-task distinction underscores the importance of compositional modeling in capturing not just the components of tasks and skills but also their contextual modifiers. In this conceptualization, tasks entail certain skills but not vice versa.

Attitudes. Other traits, often referred to as *soft skills*, are captured through entities of type ATTITUDE in our approach. We define attitudes as psychological, emotional, or behavioral predispositions—such as empathy, adaptability, or stress tolerance—that contribute to effective task performance (Rodrigues et al., 2021). Unlike skills, which are inherently linked to specific tasks, attitudes pertain to broader domains of competence.

Other entities and relations. The other entities and relations have been derived inductively during annotation guideline development (see Section 2.2) based on the goals of our framework (e.g., FORMAL QUALIFICATION was introduced because we were interested in degrees mentioned), their frequent occurrence in patterns (e.g. URGENCY) or the need to correctly represent the meaning of the text (e.g. syntactically motivated relations like COORDINATION or REP. The most arbitrary categories are ATTRIBUTES and ZERO RELATION. Attributes provide additional context that may or may not be relevant for the analysis. Attributes cannot stand alone; they specify details about primary entities. While Attributes may span longer phrases, all other entity types are defined as concisely as possible to balance annotation consistency and model performance. This design minimizes complexity for key entities while preserving optional, nuanced information through attributes, serving as a flexible "catch-all" category for contextual details. The Zero Relation on the other hand is used for related entities that require no additional specification because their connection is self-evident.

2.2 Dataset Annotation

To prepare a suitable dataset for annotation, we sampled 500 German job ads from Textkernel's Jobfeed corpus, restricting to regular employment (excluding apprenticeships). A multivariate sampling approach balanced multiple factors (year of publishing, website source, WZ08 activity, ISCO08 occupation, contract type, and text length), aiming to minimize selection bias.

We conducted the annotation in two phases: (1) iterative guideline development and (2) final annotation of 500 OJAs:

Phase 1 Following Reiter et al. (2019), four original annotators (A) refined the guidelines over six rounds on small samples, comparing annotations and adjusting rules to ensure consistency and construct validity.

Phase 2 In the final phase (2), 15 researchers (A plus newly trained annotators B) participated. Group B received tutorials and performed test annotations; only those surpassing Krippendorff's $\alpha \geq 0.7$ proceeded. Each OJA was then double-annotated and curated by a third annotator (A). This yielded Krippendorff's $\alpha = 0.88$ for entities and $\alpha = 0.80$ for relations — values considered reliable by Krippendorff (2018).

Comparing our metrics to other work in the field, Green et al. (2022) report Cohen's $\kappa = 0.49$ and Krippendorff's $\alpha = 0.55$, while Zhang et al. (2022a) report Fleiss' κ between 0.70 and 0.75. Although the scores are not directly comparable due to differences in annotation schemes and task definitions, our results indicate a relatively high inter-annotator reliability.

Resulting dataset. The annotation process resulted in a dataset of 500 German-language job ads, annotated with 22,506 entities and 13,324 relations. We refer to this dataset as GOJA ("German Online Job Advertisements").

3 Evaluation

To gain insights into the validity and usefulness of our proposed compositional annotation schema, we analyze our dataset GOJA (section 3.1) and showcase how our schema facilitates downstream analytics use cases (section 3.2). We then use GOJA to train classification models to predict our proposed annotations, present benchmark numbers and qualitatively evaluate model output (section 3.3).

3.1 Dataset Insights

Our compositional approach aims to capture complex, multi-component requirements. To assess their prevalence across GOJA, we analyze entity and relation distributions, as well as occurrences of longer compositional chains. Given our multivariate sampling approach, this distribution should approximate their occurrence in larger datasets.

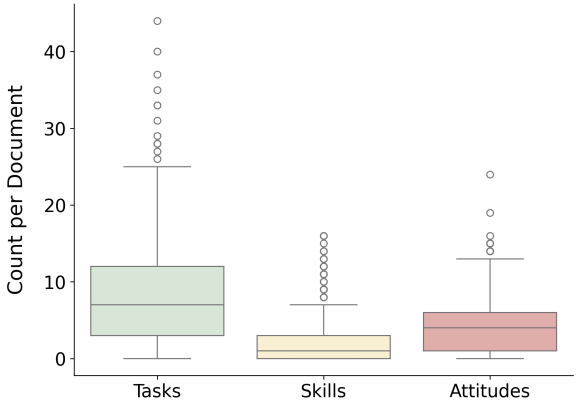


Figure 3: Boxplot showing the distributions of Skills, Tasks and Attitudes per document.

General metrics. Our dataset of 500 annotated OJAs comprises 22,506 entities and 13,324 relations, with the longest compositional chain connecting up to 36 entities. Figure 3 illustrates the distribution of key analytical units—tasks, skills, and attitudes—per document, as derived from the chains described in Section 2.1.

Explicit distinction between tasks and skills. Notably, concepts that are extracted as skills in other studies tend to be formulated as tasks in our conceptualization. This observation reflects how most analyses with OJA data (implicitly) equate job tasks with skills, i.e. the proficiency in these tasks. However, as employer-provided training is almost ubiquitous in Germany, especially in entry-level jobs in Germany (Lukowski et al., 2021), candidates are not expected to master all tasks at the outset. Consequently, our findings indicate that research could benefit from investigating why certain tasks are explicitly associated with an experience level while others are not.

Comparing the frequency of skills and attitudes, it can be derived that in terms of typical OJA text zones (Gnehm, 2018; Gnehm and Clematide, 2020), our analysis reveals that skill segments in job advertisements predominantly consist of attitudes rather than hard skills.

High frequency of conjoined skills and tasks. The analysis of conjoined and non-conjoined patterns demonstrates the widespread occurrence of complex structural patterns in OJAs. Specifically, Table 2 summarizes the frequency of various conjoined and non-conjoined patterns. 44% of tasks exhibit conjoined structures—either with a single process linked to multiple work contents or vice versa. In the case of skills, 30% of experience-level-to-task chains and 27% of experience-level-to-tool links are conjoined. These figures indicate a higher frequency of conjoined skills and tasks compared to the findings of Nguyen et al. (2024).

Pattern	Frequency
Process → Work Content	1706
Process, Process … → Work Content	609
Process → Work Content, Work Content …	707
Experience Level → Task	245
Experience Level → Task, Task …	105
Experience Level → Tool	284
Experience Level → Tool, Tool …	106

Table 2: Frequency comparison of different conjoined patterns in tasks and skills, where identical entities denote an arbitrary number n. Note that tasks in this visualization represent the aggregation of their own entity chains, combined here to maintain visual clarity.

The issue of conjoined patterns becomes even more substantial when accounting for additional relations between the entities in our analysis, as illustrated in Table 1. For instance, if an experience level is linked to an urgency indicator, this relation may need to be applied to both tasks and tools.

3.2 Exemplary Analysis

To illustrate the analytical advantages of our approach, we conducted an exemplary analysis assessing the urgency level of skills and attitudes using an mDeBERTa-based zero-shot classification model (Laurer et al., 2024). Entity pairs with an "urgency" relation were categorized into three levels: required, preferable, and unimportant. The results indicate that, when explicitly mentioned, skills and attitudes are most frequently classified as preferable (see Table 3). In some cases, they are even explicitly stated as not required, which adds particular saliency to the mentioning of this skill.

Level of Urgency	Examples	Count
Preferable	*desirable*, *advantageous*,...	312
Required	*necessary*, *mandatory*,...	127
Unimportant	*not necessary*, *not required*,...	27

Table 3: Overview of urgency classifications for skills and attitudes based on a zero-shot classification model. The table lists example terms associated with each urgency level and their frequency in the dataset.

Model	**Entity Extraction (F1)**	**Relation Classification (F1)**
German BERT	0.665 ± 0.025	0.836 ± 0.008
German DistilBERT	0.517 ± 0.024	0.788 ± 0.012
jobBERT-de	0.718 ± 0.013	0.874 ± 0.014
XLM-RoBERTa	**0.856** ± 0.012	**0.911** ± 0.007

Table 4: F1 scores and standard deviation for entity extraction and relation classification, averaged over five-fold cross-validation.

This distinction is, for example, relevant for OJA-to-CV matching systems, as Fazel-Zarandi and Fox (2009) specifically differentiate between must-have and nice-to-have skills in their work.

Another example of the analytical advantages of our approach is the classification of alternative relations. Our dataset shows that in roughly 5% of job ads, formal qualifications (e.g., degrees, vocational training) could be substituted if the candidate had sufficient work experience. These examples demonstrate that, on the one hand, considering requirements in OJAs without their context can lead to misleading results, and on the other hand, our detailed modeling offers valuable potential for both practice and research.

3.3 Model Benchmark and Evaluation

In this section, we train and evaluate models on our dataset to establish baseline performance numbers and gain further insights into the validity of our approach.

3.3.1 Experimental Setup

We fine-tune four different pre-trained transformer models: German BERT (Devlin et al., 2019), German DistilBERT (Sanh, 2019), jobBERT-de (Gnehm et al., 2022b)—a variant of German BERT fine-tuned on German OJA data—and the multilingual XLM-RoBERTa (Conneau, 2019). For entity extraction, we use a token classification head on top of the pre-trained models.

For relation classification, we adopt a simple yet effective approach: Entities participating in a relation are marked with special tokens [E] and [/E] within their sentence, and the modified sequence is passed to a transformer-based sequence classification model. To handle candidate entity selection efficiently, we use a context window of four sentences, based on internal analyses, to determine potential entity pairs. Additionally, we introduce a NO RELATION class to distinguish entity pairs that do not share a relation. Since this results in a class imbalance, we randomly downsample the No Relation class to match the total number of instances in the other relation classes.

Prior to cross-validation, we determine suitable hyperparameters via grid search to optimize model performance. We report the F1-score averaged over five-fold cross-validation, ensuring robustness across different data splits. The dataset follows a 70-15-15 split into training, validation, and test sets, with all reported F1-scores computed exclusively on the unseen test set to provide a realistic assessment of generalization performance.

3.3.2 Experimental Results

Our experimental results are summarized in Table 4. We observe that XLM-RoBERTa clearly outperforms the other three models in both entity extraction and relation classification. Notably, jobBERT-de also achieves solid performance, improving over German BERT and German DistilBERT in both tasks. An interesting finding is that the performance gap among models is much larger in the entity extraction subtask than in relation classification.

3.3.3 Error Analysis

Our error analysis aims to explain model performance differences on a per-class level and to understand the relationship between model predictions, inter-annotator agreement (IAA), and error patterns. Figure 4 presents per-class F1 scores and std. deviations, while confusion matrices (Figures 6 and 7) illustrate detailed prediction errors. Our analysis shows that superior macro-F1 scores of XLM-RoBERTa stem primarily from its ability to handle difficult classes rather than from general peak performance.

Weak classes. Entity extraction errors cluster around three difficult classes: Formal Qualification (FQ), Attribute, and Industry. Relation extraction errors are concentrated in Degree of Autonomy and REP. Attribute and Industry are conceptually difficult, reflected in low IAA scores. Attribute

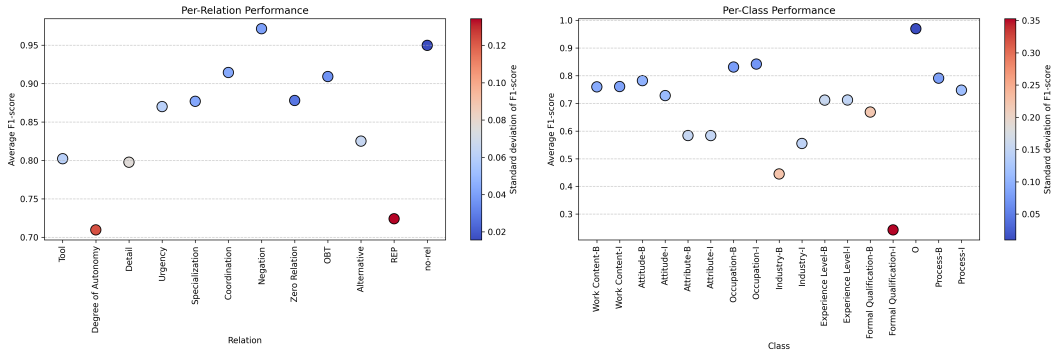


Figure 4: Mean F1-score across all models for each entity and relation class. The color gradient represents the standard deviation of F1-scores across runs.

acts as a broad, catch-all category with long and inconsistent spans, while Industry annotations are limited to candidate-focused sections, causing ambiguity about what qualifies as an industry mention. Both classes are frequently confused with the Outside (O) label, as shown in the confusion matrices, which is less critical since these errors often reflect borderline cases rather than clear misclassifications.

A similar pattern appears in relation classification: Degree of Autonomy and REP have low IAA scores and few examples, resulting in low F1 scores. In contrast, other classes with low IAA scores, such as Zero Relation and Specialization, perform better due to having more examples. The high performance of Negation, despite having few examples, further suggests that performance depends on both conceptual clarity and class frequency.

FQ as a notable outlier. Although the FQ class exhibits high IAA scores and clear conceptual boundaries, it performs poorly for all models except XLM-RoBERTa. Confusion matrices reveal that weaker models seldom predict `FQ-I` at all. Besides the general overprediction of the outside class, the models show different behavior in regard to FQ. DistilBERT models frequently predict Work Content-I, Attribute-I, or Experience Level-I instead of FQ-I. Manual inspection shows that these models often switch from FQ-B to the inside tag of another entity type mid-span. Both the internal splitting of spans and the confusion between semantically distinct entity types are notable and unexpected. In contrast, BERT and jobBERT-de models display a different error pattern: they tend to predict FQ-B but fail to continue the span with FQ-I, predicting another FQ-B. Only XLM-RoBERTa is able to predict FQ reliably.

4 Discussion

Our analyses confirm the effectiveness of our compositional modeling approach. This is supported by strong IAA scores and solid XLM-RoBERTa performance, demonstrating that high-quality training data is achievable despite our more detailed and comprehensive semantic modeling. Moreover, our data analysis shows that complex patterns, such as conjoined skills, are not rare edge cases but occur frequently in OJAs, underscoring the analytical advantages of our approach for capturing such structures.

Model comparison. We note that there is only limited possibility to compare our results with existing work, since we propose a shift from a pure span-based classification task to a combined entity and relation extraction approach. Furthermore, our approach captures multiple requirement types, unlike prior work focusing on isolated concepts.

Additionally, Zhang et al. (2023) show that skill extraction performance varies widely across datasets, with Span-F1 scores ranging from 45.6% to 92.2%. This variability highlights that extraction performance heavily depends on how skills are modeled, with simpler formulations often achieving higher scores but lacking structural and semantic depth (see also Alexopoulos, 2020).

Emerging compositional cpproaches in OJA research. While our method focuses on capturing intricate structures through entity- and relation-level modeling, other studies have introduced alternative solutions to similar challenges. Nguyen et al. (2024) reframe skill extraction from a BIO sequence labeling task to a large language model (LLM) generation task, aiming to better capture complex patterns and reduce labeled data requirements. While their approach improves generaliz-

ability, our method models logical and semantic relationships explicitly. Moreover, their error analysis highlights that both traditional and generative models struggle with conjoined skills—a challenge our compositional framework handles more effectively. Additionally, encoder-based models are typically more computationally efficient than large generative models, making them more practical for large-scale applications.

Gnehm et al. (2022a) pursue a two-step approach to fine-grained skill extraction, first identifying long skill spans containing multiple subcomponents and then classifying these subcomponents—some of which align with our entities types, such as their "container" concept, which to some extend corresponds our Experience Level entity. While their method also adopts a compositional perspective, we argue that our combined token and relation classification approach is more flexible and efficient. Specifically, our framework can selectively ignore irrelevant details within lengthy spans, while their method may struggle with overly long initial spans.

Broader applicability. Compositional modeling of entities and concepts is not unique to our approach; it also underlies many relation extraction tasks where relations between entities construct higher-order concepts. While traditional relation extraction typically operates on classic named entities, our method starts from predefined conceptual structures and decomposes them into text-based components. Despite differences in granularity, both approaches transform lower-level units into more complex representations.

Beyond standard relation extraction, compositional modeling also appears in tasks that derive non-discrete concepts from interrelated textual elements. Event detection constructs events from triggers and arguments (Xiang and Wang, 2019), while Song et al. (2021) apply relation extraction to verb metaphors, capturing related entities without fully decomposing the metaphor itself. Similarly, aspect-based sentiment analysis (ABSA) (Zhang et al., 2022b) assigns sentiments to specific text spans rather than entire documents, applying the principle of decomposition on a textual level, whereas our approach decomposes entities into smaller subentities and models their relations.

We believe that the broader NLP community, particularly in application-driven fields such as industry, computational social science (CSS), and digital humanities (DH), could benefit from a more extensive discussion on compositionality in text and its relation to conceptual modeling. Our findings highlight the limitations of treating many information extraction IE tasks purely as named entity recognition (NER) problems.

5 Conclusion and Outlook

This paper introduced a compositional entity modeling framework for requirement extraction from Online Job Advertisements (OJAs). Our approach reframes the task from token classification to a combined entity and relation classification task, effectively capturing complex, multi-component requirement structures. Through our gold-standard dataset of 500 annotated German job ads, we demonstrated high annotation consistency and outperformance of traditional span-based approaches in capturing critical semantic dependencies.

Our work leaves room for future research. While our dataset and experiments focus on German OJAs, future research should investigate whether the compositional modeling approach yields similar benefits for other languages. Additionally, future research should adopt a more extensive benchmarking process that integrates additional evaluation metrics—such as graph-based measures (triples)—the aggregation of target concepts like tasks and skills, and advanced model architectures (e.g., joint entity-relation extraction models Shaowei et al., 2022 or graph-based neural networks Wu et al., 2020).

Beyond extraction, requirement analysis often involves mapping extracted entities to taxonomies or ontologies. Since ontologies can be represented as graphs (see Dörpinghaus et al., 2023 for a relevant ontology for our dataset), our triplet-based extraction approach, which produces graph-like structures, may facilitate joint or hierarchical taxonomy alignment. Moreover, evaluating the benefits of compositional modeling in downstream applications, such as job-to-candidate matching or labor market analyses, could highlight its practical impact.

In conclusion, our compositional modeling framework contributes to a deeper understanding of complex entities in applied NLP tasks and lays the groundwork for future innovations in requirement extraction and structured entity modeling.

6 Limitations

While our compositional entity modeling framework shows promising results in capturing complex semantic dependencies in online job advertisements, several limitations and deliberate design decisions should be acknowledged.

Limited Large-Scale Empirical Validation. Although our experiments indicate that the proposed method can more effectively capture the intricate structure of job requirements compared to traditional single-span extraction methods, conclusively validating this claim would require large-scale empirical comparisons across diverse modeling paradigms. Such an endeavor would involve developing and benchmarking multiple models on datasets comprising millions of OJAs and assessing their performance across various downstream applications (e.g., skill gap analysis, regional labor market assessments). Given the substantial scope and resource requirements, this comprehensive evaluation remains beyond the scope of the current study.

Design Decisions in Entity and Relation Definitions. A central design choice of our framework is to consistently label similar textual components with the same entity type—specifically, using work content for elements that denote the object or subject within a sentence. For example, a machine mentioned in a job advertisement is always annotated as Work content, irrespective of whether the context involves repairing or operating machinery. The semantic differences between these contexts are then captured through distinct relation types: when the machine is directly acted upon (as in repairing machinery), the relation OBT is used, whereas if it serves as an instrument (as in operating machinery), the relation Tool is applied. This choice was made, because we believe it would enhance annotation consistency and model performance.

Then, other relational distinctions, such as Alternative, emerge directly from the logical structure of the text. However, decisions regarding when to introduce a new entity versus representing semantic nuances solely through relations (e.g., the case of Specialization, which often maps to attributes) proved challenging and, in some cases, inherently arbitrary. These design choices could affect both the generalizability of the framework and the interpretability of the extracted structures. Balancing the need for annotation consistency with the capture of fine-grained semantic distinctions remains an open challenge and a potential limitation of our approach.

Context Window and Sentence Splitting. For relation classification, we sample candidate entity pairs within a context window defined by sentence boundaries. This decision was based on analyses suggesting that sentences provide a natural and less arbitrary segmentation unit compared to tokens or words. However, sentence splitting in job advertisements is challenging due to unconventional punctuation, enumerations, and gender-neutral formulations in German. Such issues can lead to suboptimal context sizes, potentially affecting the capture of relevant relational dependencies. Future work should investigate more robust segmentation strategies.

Token Alignment Issues. Our annotations are performed at the character level and subsequently aligned with tokenized text. In rare cases, discrepancies between token boundaries and annotated spans occur. Although internal analysis indicates that these misalignments are marginal, they nonetheless represent a potential source of error that might slightly affect extraction performance during inference. Addressing these alignment challenges is an important direction for future research. Note, that this problem did not affect the model performances presented in Section 3.3.

Comparison with Single-Span Extraction Approaches. A potential counterargument is that extracting entire spans in a single step might allow for the resolution of semantic and logical connections in later processing stages. However, research (Zhang et al., 2022a) has shown that longer spans become increasingly difficult for models to extract accurately. Thus, while a single-span approach might postpone the need to capture internal structure, it does not eliminate the inherent challenges associated with modeling complex semantic relationships in job advertisements.

References

Panos Alexopoulos. 2020. *Semantic Modeling for Data*. O'Reilly Media.

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2018. New technologies and the labor market. *Journal of Monetary Economics*, 97:48–67.

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2020. The evolution of work in the united states. *American Economic Journal: Applied Economics*, 12(2):1–34.

Enghin Atalay, Sebastian Sotelo, and Daniel Tannenbaum. 2023. The geography of job tasks. *Journal of Labor Economics*.

David H Autor and Michael J Handel. 2013. Putting tasks to the test: Human capital, job tasks, and wages. *Journal of labor Economics*, 31(S1):S59–S96.

Rahkakavee Baskaran and Johannes Müller. 2023. Classification of german job titles in online job postings using the kldb2010 taxonomy. Last accessed: 2024-05-22.

Adam SZ Belloum, Spiros Koulouzis, Tomasz Wiktorski, and Andrea Manieri. 2019. Bridging the demand and the offer in data science. *Concurrency and Computation: Practice and Experience*, 31(17):e5200.

Katy Börner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James A. Evans. 2018. Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences of the United States of America*, 115(50):12630–12637.

Phillip Brown and Manuel Souto-Otero. 2020. The end of the credential society? an analysis of the relationship between education and the labour market using big data. *Journal of Education Policy*, 35(1):95–118.

Marlis Buchmann, Helen Buchs, Felix Busch, Simon Clematide, Ann-Sophie Gnehm, and Jan Müller. 2022. Swiss job market monitor: A rich source of demand-side micro data of the labour market. *European Sociological Review*, 38(6):1001–1014.

Emilio J Castilla and Hye Jin Rho. 2023. The gendering of job postings in the online recruitment process. *Management Science*, 69(11):6912–6939.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Jens Dörpinghaus, Johanna Binnewitt, Stefan Winnige, Kristine Hein, and Kai Krüger. 2023. Towards a german labor market ontology: Challenges and applications. *Applied Ontology*, (Preprint):1–23.

Marta Fana, Martina Bisello, Sergio Torrejón Pérez, and Enrique Fernández-Macías. 2023. What workers do and how. a european database of tasks indices.

Maryam Fazel-Zarandi and Mark S Fox. 2009. Semantic matchmaking for job recruitment: an ontology-based hybrid approach. In *Proceedings of the 8th International Semantic Web Conference*, volume 525, page 2009.

Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. 2021. Neo: A system for identifying new emerging occupation from job ads. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18):16035–16037.

Ann-Sophie Gnehm. 2018. Text zoning for job advertisements with bidirectional lstms. In *Proceedings of the 3rd Swiss Text Analytics Conference (SwissText 2018)*, pages 1–9, Winterthur. University of Zurich.

Ann-sophie Gnehm, Eva Bühlmann, Helen Buchs, and Simon Clematide. 2022a. Fine-grained extraction and classification of skill requirements in german-speaking job ads. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*. Association for Computational Linguistics.

Ann-Sophie Gnehm, Eva Bühlmann, and Simon Clematide. 2022b. Evaluation of transfer learning and domain adaptation for analyzing german-speaking job advertisements. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Ann-Sophie Gnehm and Simon Clematide. 2020. Text zoning and classification for job advertisements in german, french and english. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 83–93.

Thomas AF Green, Diana Maynard, and Chenghua Lin. 2022. Development of a benchmark corpus to support entity recognition in job descriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1201–1208. European Language Resources Association.

Betül Güntürk-Kuhl, Philipp Martin, and Anna Cristin Lewalder. Die taxonomie der arbeitsmittel des bibb: Revision 2018.

Imane Khaouja, Ismail Kassou, and Mounir Ghogho. 2021. A survey on skill identification from online job ads. *IEEE Access*, 9:118134–118153.

Jeonghyun Kim and Putthachat Angnakoon. 2016. Research using job advertisements: A methodological assessment. *Library & Information Science Research*, 38(4):327–335.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Nan Li, Bo Kang, and Tijl de Bie. Llm4jobs: Unsupervised occupation extraction and standardization leveraging large language models.

Antonio Lima, B Bakhshi, et al. 2018. Classifying occupations using web-based job advertisements: an application to stem and creative occupations. *Economic Statistics Centre of Excellence Discussion Paper*, 8.

Felix Lukowski, Myriam Baum, and Sabine Mohr. 2021. Technology, tasks and training–evidence on the provision of employer-provided training in times of technological change in germany. *Studies in Continuing Education*, 43(2):174–195.

Johannes Müller. Machbarkeitsstudie: Teilqualifikationen in online-job-anzeigen (oja): Methodenbericht zur automatisierten extraktion von teilqualifikationen für fünf ausbildungsberufe: Projekt: Aufstieg durch kompetenzen.

Khanh Nguyen, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. 2024. Rethinking Skill Extraction in the Job Market Domain using Large Language Models. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 27–42, St. Julian's, Malta. Association for Computational Linguistics.

Dimos Ntioudis, Panagiota Masa, Anastasios Karakostas, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2022. Ontology-based personalized job recommendation framework for migrants and refugees. *Big Data and Cognitive Computing*, 6(4):120.

Ibrahim Rahhal, Kathleen M. Carley, Ismail Kassou, and Mounir Ghogho. 2023. Two stage job title identification system for online job advertisements. *IEEE Access*, 11:19073–19092.

Nils Reiter, Marcus Willand, and Evelyn Gius. 2019. A shared task for the digital humanities chapter 1: Introduction to annotation, narrative levels and shared tasks. *Journal of Cultural Analytics*, 4(3).

Margarida Rodrigues, Fernández-Macías, and Enrique, Sostero, Matteo. 2021. A unified conceptual framework of tasks, skills and competences.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Benjamin Schimke. 2023. Nachweise für berufliche qualifikationen oder doch nur ein motivationssignal? zur wirkung non-formaler weiterbildungszertifikate in der personalauswahl. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 75(4):451–475.

Elena Senger, Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings. *Preprint*, arXiv:2402.05617.

ZHANG Shaowei, WANG Xin, CHEN Zirui, WANG Lin, XU Dawei, and JIA Yongzhe. 2022. Survey of supervised joint entity relation extraction methods. *Journal of Frontiers of Computer Science & Technology*, 16(4).

Ellery Smith, Andreas Weiler, and Martin Braschler. 2021. Skill extraction for domain-specific text retrieval in a job-matching platform. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 116–128. Springer.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4240–4251.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Mike Zhang, Kristian Nørgaard Jensen, Sif Dam Sonniks, and Barbara Plank. 2022a. Skillspan: Hard and soft skill extraction from english job postings. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Mike Zhang, Rob van der Goot, and Barbara Plank. 2023. ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11871–11890, Toronto, Canada. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022b. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

A Additional examples

Figure 5 provides the full sentence of the example in Section 1, including different annotation variants of traditional approaches and our solution.

B Dataset Details

Data Sampling. To reduce biases, for example due to data shift or OJAs differing between jobs or industry sectors, we applied a multivariate sampling approach. Table 5 explains the different variables used.

Annotation guidelines. Annotation guidelines can be accessed under https://github.com/TM4VETR/Public_Stea_Annotationsguide

Annotators. All annotators (A+B) work in the same organization as the authors of this article. They are all native German speakers and hold at least the equivalent of a Bachelor's degree, with diverse backgrounds in social sciences, (digital) humanities, economics, and psychology. All have at least some experience in labor market research, which is advantageous given the complex structure of the operationalization of the concepts. Four of the annotators are male, and eleven are female.

All annotations were conducted during regular working hours, and the annotators did not receive any additional payment beyond their regular salary. All B participated voluntarily following a call for participation.

The annotators were informed about the purpose of the annotation process, and in exchange for their contribution, they were promised priority access to the final dataset.

Additional IAA scores. Tables 6 and 7 show the IAA results per class.

Entity and relation counts. Table 8 displays of the amount of annotated entities and relations in our dataset.

C Experimental Setup Details

To ensure reproducibility, we provide additional details on our experimental setup:

Hyperparameters. Table 9 and **??** provide details regarding the hyperparameters used in our experiments.

Hardware: All models were trained on an NVIDIA L40 GPU with 48 GB VRAM.

Class Imbalance: The "No Relation" class was downsampled to match the total number of instances in other relation classes.

Cross-Validation: A stratified 5-fold cross-validation was performed using the same five random seeds across all models.

Licences:

D Additional Analysis

Figures 6 and 7 display the aggregated confusion matrices for entity extraction and relation classification, respectively, across five runs per model. As they do use numeric labels for space reasons, the label mapping presented in Tables 11 and 12 respectively.

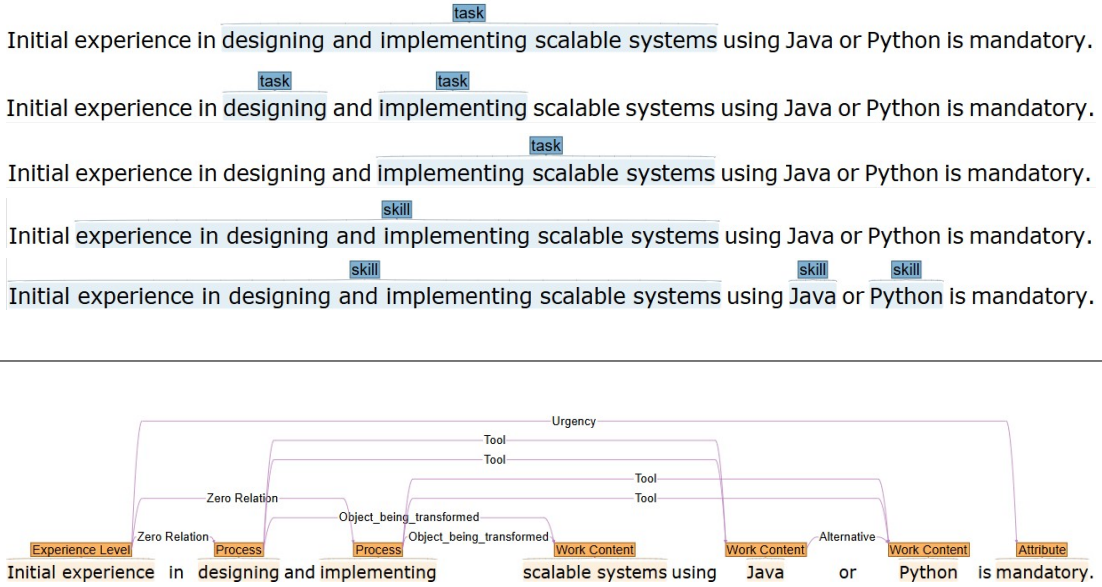
E Information About Use Of AI Assistants

We used AI assistants as a tool to support both the writing and coding aspects of this research. In particular, AI-assisted tools were employed to generate initial drafts of text, suggest improvements in language and structure, and assist with coding tasks. All AI-generated content was thoroughly reviewed, refined, and integrated by the authors to ensure accuracy, clarity, and alignment with our research objectives. The use of AI was solely aimed at increasing efficiency in routine tasks, and final decisions and edits were made by the research team.

F Ethics statement

Our study is purely academic in nature, and we do not foresee any significant risks or adverse impacts arising from our approach. The dataset used consists of non-public job advertisements and has been processed strictly for research purposes, with all sensitive information anonymized prior to analysis. Given that our methodology is applied solely for analytical and evaluation objectives, we believe that our work does not pose any harm.

Figure 5: Extended Example: Comparison of Traditional Approaches vs. Our Approach



Above: Traditional Extraction Methods. Below: Our Compositional Approach.

Factor	**Description**
Year of Publishing	Job ads from the years 2016 and 2022.
Source Website	Job portals and company websites.
WZ08 Activity	Selection from the economic sections of the WZ08 classification.
ISCO08 Occupation	First level of the ISCO08 occupational classification.
Contract Type	Only permanent and fixed-term contracts (excluding apprenticeships, internships, etc.).
Text Length	Various text lengths, measured using spaCy tokenization.

Table 5: Factors in the Multivariate Sampling Approach for Job Ad Selection

Entity Type	**IAA (Krippendorff's α)**
Work Content	0.75
Attitude	0.87
Attribute	0.60
Occupation	0.83
Industry	0.55
Experience Level	0.85
Formal Qualification	0.87
Process	0.78

Table 6: Inter-Annotator Agreement (Krippendorff's α) for Entity Types

Relation Type	**IAA (Krippendorff's α)**
Alternative	0.75
Coordination	0.75
Degree of Autonomy	0.62
Detail	0.62
Negation	0.90
Object Being Transformed (OBT)	0.72
Related Entity Parts (REP)	0.67
Specialization	0.68
Tool	0.61
Urgency	0.78
Zero Relation	0.52

Table 7: Inter-Annotator Agreement (Krippendorff's α) for Relation Types

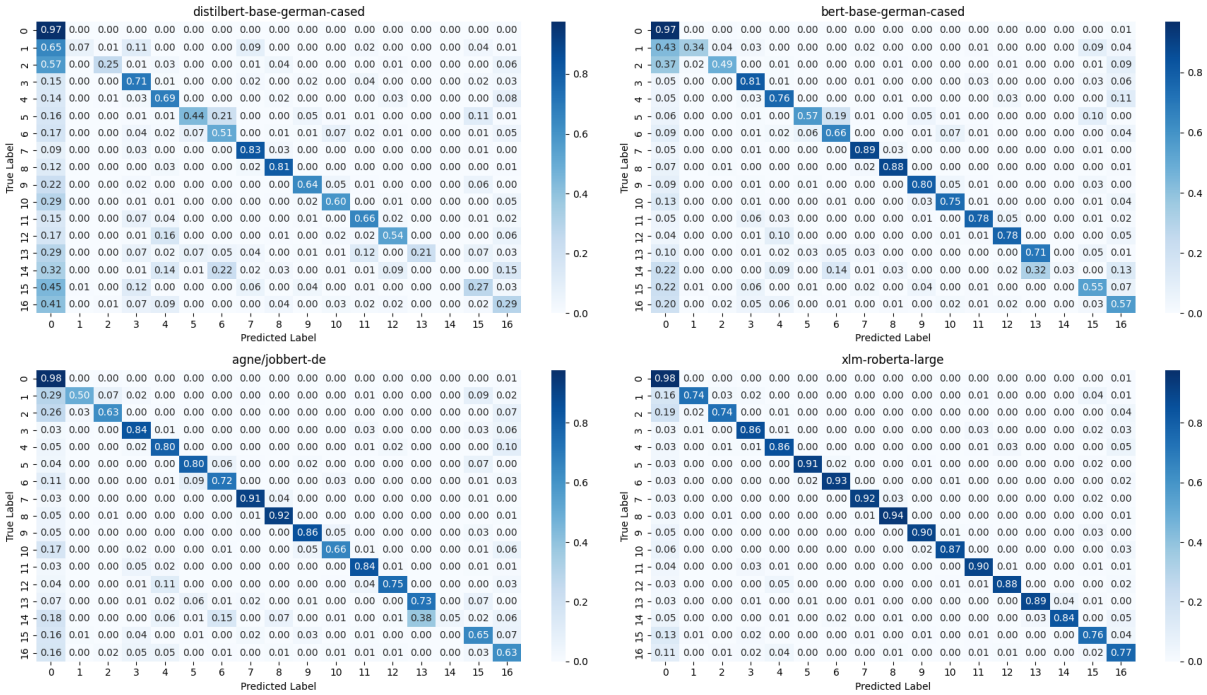


Figure 6: Aggregated confusion matrices for entity extraction (row-normalized over 5 runs for each model)

Entities	**Count**
Work Content	5285
Attribute	4685
Process	4461
Attitude	2172
Occupation	2105
Industry	1615
Experience Level	1412
Formal Qualification	771
Relations	**Count**
Zero Relation	4322
OBT	3648
Specialization	1345
Tool	1157
Alternative	597
Detail	585
Coordination	482
Urgency	466
Degree of Autonomy	325
REP	312
Negation	85

Table 8: Number of annotated entities and relations per class

Task	XLM-RoBERTa	jobBERT-de, German BERT	DistilBERT
Entity Extraction	7 epochs	9 epochs	15 epochs
Relation Classification	6 epochs	8 epochs	12 epochs

Table 9: Number of epochs per model

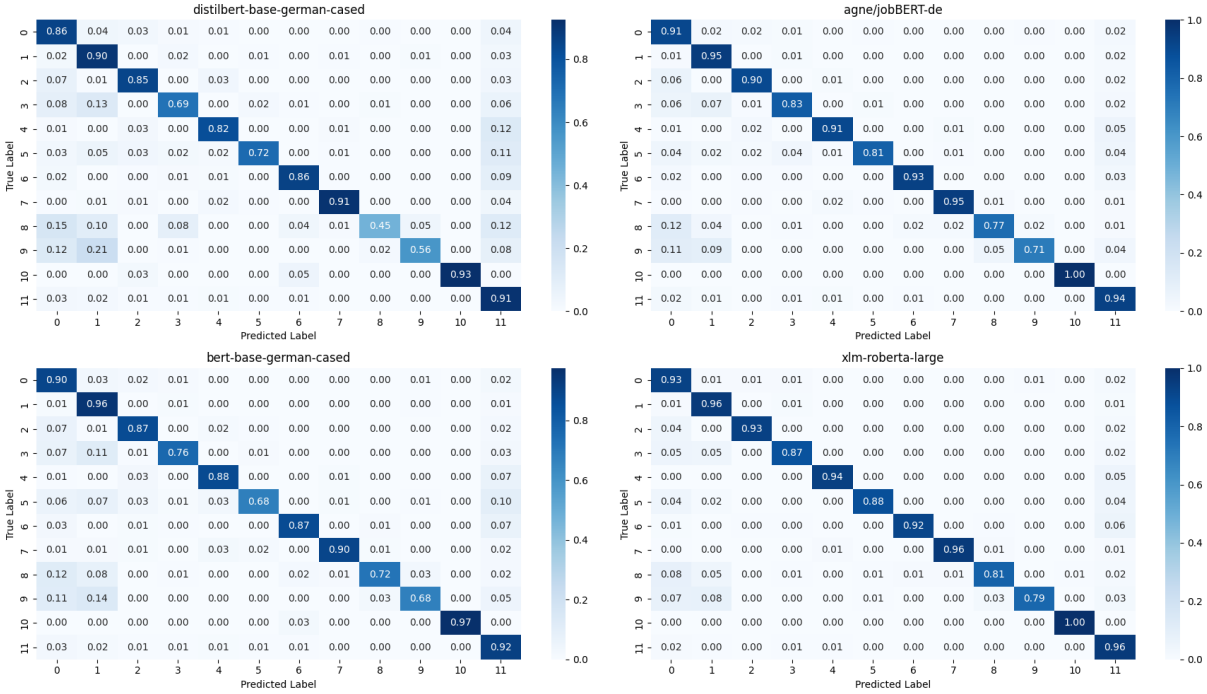


Figure 7: Aggregated confusion matrices for relation classification (row-normalized over 5 runs for each model)

Hyperparameter	Value
Batch Size	64 (XLM-RoBERTa: 16)
Learning Rate	5e-5
Weight Decay	0
Adam Betas	(0.9, 0.999)
Adam Epsilon	1e-8
Max Gradient Norm	1.0
Scheduler	Linear
Warmup Ratio	0.0

Table 10: Hyperparameter details

Model	License
MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c	MIT License
google-bert/bert-base-german-cased	MIT License
distilbert/distilbert-base-german-cased	Apache License 2.0
agne/jobBERT-de	CC-BY-NC-SA 4.0
FacebookAI/xlm-roberta-base	MIT License

Label number	Label name
0	O
1	Industry-B
2	Industry-I
3	Work Content-B
4	Work Content-I
5	Experience Level-B
6	Experience Level-I
7	Occupation-B
8	Occupation-I
9	Attitude-B
10	Attitude-I
11	Process-B
12	Process-I
13	Formal Qualification-B
14	Formal Qualification-I
15	Attribute-B
16	Attribute-I

Table 11: Entity label mapping

Label number	Label number
0	Zero Relation
1	OBT
2	Specialization
3	Tool
4	Alternative
5	Detail
6	Urgency
7	Coordination
8	REP
9	Degree of Autonomy
10	Negation
11	no-rel

Table 12: Relation label mapping