Skill Extraction from Job Postings using Weak Supervision

Mike Zhang[1,*], Kristian Nørgaard Jensen[1], Rob van der Goot[1] and Barbara Plank[1,2]

[1]IT University of Copenhagen, Rued Langgaards Vej 7, 2300, Copenhagen, Denmark
[2]Ludwig Maximilian University of Munich, Akademiestraße 7, 80799, Munich, Germany

Abstract
Aggregated data obtained from job postings provide powerful insights into labor market demands, and emerging skills, and aid job matching. However, most extraction approaches are supervised and thus need costly and time-consuming annotation. To overcome this, we propose Skill Extraction with Weak Supervision. We leverage the European Skills, Competences, Qualifications and Occupations taxonomy to find similar skills in job ads via latent representations. The method shows a strong positive signal, outperforming baselines based on token-level and syntactic patterns.

Keywords
Skill Extraction, Weak Supervision, Information Extraction, Job Postings, Skill Taxonomy, ESCO

1. Introduction

The labor market is under constant development—often due to changes in technology, migration, and digitization—and so are the skill sets required [1, 2]. Consequentially, large quantities of job vacancy data is emerging on a variety of platforms. Insights from this data on labor market skill set demands could aid, for instance, job matching [3]. The task of automatic *skill extraction* (SE) is to extract the competences necessary for any occupation from unstructured text.

Previous work on supervised SE frame it as a sequence labeling task (e.g., [4, 5, 6, 7, 8, 9, 10]) or multi-label classification [11]. Annotation is a costly and time-consuming process with little annotation guidelines to work with. This could be alleviated by using predefined skill inventories.

In this work, we approach span-level SE with weak supervision: We leverage the European Skills, Competences, Qualifications and Occupations (ESCO; [12]) taxonomy and find similar spans that relate to ESCO skills in embedding space (Figure 1). The advantages are twofold: First, labeling skills becomes obsolete, which mitigates the cumbersome process of annotation. Second, by extracting skill phrases, this could possibly enrich skill inventories (e.g., ESCO) by finding paraphrases of existing skills. We seek to answer: *How viable is Weak Supervision in the context of SE?* We contribute: ① A novel weakly supervised method for SE; ② A linguistic analysis of ESCO skills and their presence in job postings; ③ An empirical analysis of different embedding pooling methods for SE for two skill-based datasets.[1]

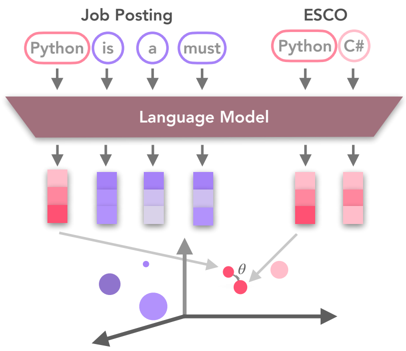

Figure 1: Weakly Supervised Skill Extraction. All ESCO skills and n-grams are extracted and embedded through a language model, e.g., RoBERTa [13], to get representations. We label *spans* from job postings close in vector space to the ESCO skill.

2. Methodology

Formally, we consider a set of job postings $\mathcal{D}$, where $d \in \mathcal{D}$ is a set of sequences (e.g., job posting sentences) with the i^{th} input sequence $\mathcal{T}_d^i = [t_1, t_2, ..., t_n]$ and a target sequence of BIO-labels $\mathcal{Y}_d^i = [y_1, y_2, ..., y_n]$ (e.g., “B-SKILL”, “I-SKILL”, “O”).[2] The goal is to use an algorithm, which predicts skill spans by assigning an output label sequence $\mathcal{Y}_d^i$ for each token sequence $\mathcal{T}_d^i$ from a job posting based on representational similarity of a span to any skill in ESCO.

RecSys in HR'22: The 2nd Workshop on Recommender Systems for Human Resources, in conjunction with the 16th ACM Conference on Recommender Systems, September 18–23, 2022, Seattle, USA.
*Corresponding author.
mikz@itu.dk (M. Zhang); krnj@itu.dk (K. N. Jensen); robv@itu.dk (R. van der Goot); b.plank@lmu.de (B. Plank)
https://jjzha.github.io/ (M. Zhang); http://kris927b.github.io/ (K. N. Jensen); http://robvanderg.github.io/ (R. van der Goot); http://bplank.github.io/ (B. Plank)

[1]https://github.com/jjzha/skill-extraction-weak-supervision
[2]Definition of labels can be found in [8].

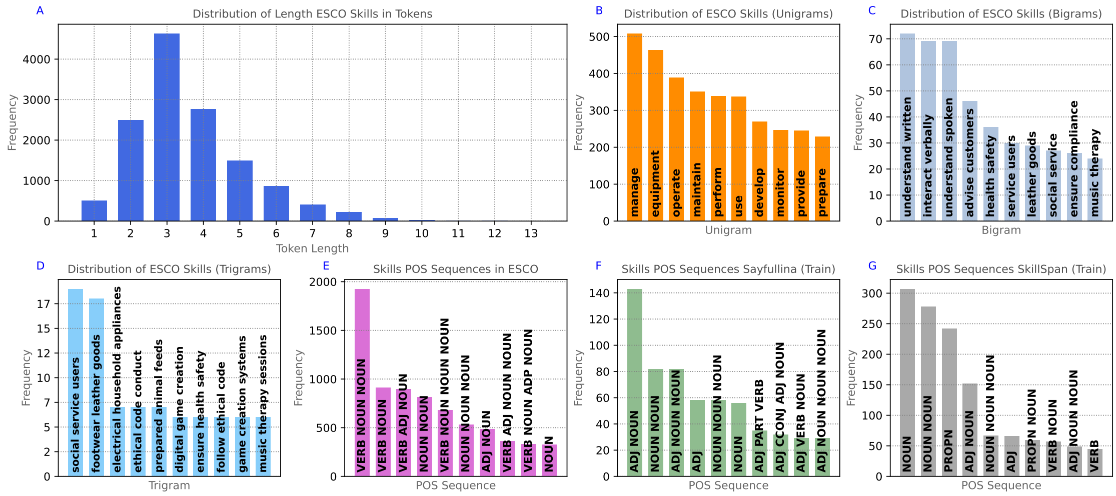

Figure 2: Surface-level Statistics of ESCO. We show various statistics of ESCO. (**A**) ESCO skills token length, the mode is three tokens. (**B**) Most frequent unigrams of ESCO skills. (**C**) Most frequent bigrams of ESCO skills. (**D**) Most frequent trigrams of ESCO skills. (**E**) Most frequent POS sequences of ESCO skills. Last, we show the POS sequences of unique skills in both train sets of *Sayfullina* and *SkillSpan* (**F**-**G**).

Table 1
Statistics of Datasets. Indicated is each dataset and their respective number of sentences, tokens, skill spans, and the average length of skills in tokens.

	Statistics	*Sayfullina*	*SkillSpan*
Train	**# Sentences**	3,703	5,866
	# Tokens	53,095	122,608
	# Skill Spans	3,703	3,325
Dev.	**# Sentences**	1,856	3,992
	# Tokens	26,519	52,084
	# Skill Spans	1,856	2,697
Test	**# Sentences**	1,848	4,680
	# Tokens	26,569	57,528
	# Skill Spans	1,848	3,093
	Avg. Len. Skills	1.77	2.92

2.1. Data

We use the datasets from [8] (*SkillSpan*) and a modification of [4] (*Sayfullina*).[3] In Table 1, we show the statistics of both. *SkillSpan* contain nested labels for skill and knowledge components [12]. To make it fit for our weak supervision approach, we simplify their dataset by considering both skills and knowledge labels as one label (i.e., `B-KNOWLEDGE` becomes `B-SKILL`).

[3]In contrast to *SkillSpan*, *Sayfullina* has a skill in every sentence, where they focus on categorizing sentences for soft skills.

ESCO Statistics We use ESCO as a weak supervision signal for discovering skills in job postings. There are 13,890 ESCO skills.[4] In Figure 2, we show statistics of the taxonomy: (A) On average most skills are 3 tokens long. In (C-D), we show n-grams frequencies with range [1; 3]. We can see that the most frequent uni- and bigrams are verbs, while the most frequent trigrams consist of nouns.

Additionally, we show an analysis of ESCO skills from a linguistic perspective. We tag the training data using the publicly available MaChAmp v0.2 model [14] trained on all Universal Dependencies 2.7 treebanks [15].[5] Then, we count the most frequent Part-of-Speech (POS) tags in all sources of data (E-G). ESCO's most frequent tag sequences are `VERB-NOUN`, these are not as frequent in *Sayfullina* nor *SkillSpan*. *Sayfullina* mostly consists of adjectives, which is attributed to the categorization of soft skills. *SkillSpan* mostly consists of `NOUN` sequences. Overall, we observe most skills consist of verb and noun phrases.

2.2. Baselines

As our approach is to find similar n-grams based on ESCO skills, we choose an n-gram range of [1; 4] (where 4 is the median) derived from Figure 2 (A). For higher matching probability, we apply an additional pre-processing step to the ESCO skills by removing non-tokens (e.g., brackets)

[4]Per 25-03-2022, taking ESCO `v1.0.9`.
[5]A Udify-based [16] multi-task model for POS, lemmatization, dependency parsing, built on top of the `transformers` library [17], and specifically using mBERT [18].

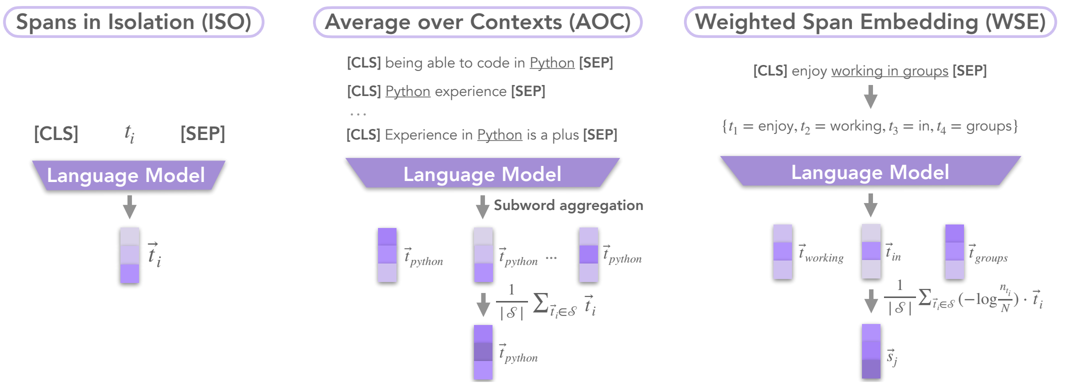

Figure 3: Skill Representations. We show different methods to embed ESCO skill phrases. The approaches are inspired by Litschko et al. [19]. We embed a skill by encoding it directly without surrounding context (**left**). We aggregate different contextual representations of the same skill term (**middle**). Last, we encode the skill phrase via a weighted sum of embeddings with each token's inverse document frequency as weight (**right**). For the middle and right methods, $\mathcal{S}$ is the number of sentences where the ESCO skill appears.

and words between brackets (e.g., "Java (programming)" becomes "Java"). We have three baselines:

Exact Match: We do exact substring matching with ESCO and the sentences in both datasets.

Lemmatized Match: ESCO skills are written in the infinitive form. We take the same approach as exact match on the training sets, now with the lemmatized data of both. The data is lemmatized with MaChAmp v0.2 [14].

POS Sequence Match: Motivated by the observation that certain POS sequences often overlap between sources (Figure 2, E-G), we attempt to match POS sequences within ESCO with the POS sequences in the datasets. For example `NOUN-NOUN`, `NOUN`, `VERB-NOUN` and `ADJ-NOUN` sequences are commonly occurring in all three sources.

2.3. Skill Representations

We investigate several encoding strategies to match n-gram representations to embedded ESCO skills, the approaches are inspired by Litschko et al. [19], where they applied them to Information Retrieval. The language models (LMs) used to encode the data are RoBERTa [13] and the domain-specific JobBERT [8]. All obtained vector representations of skill phrases with the three previous encoding methods are compared pairwise with each n-gram created from *Sayfullina* and *SkillSpan*. An explanation of the methods (see Figure 3):

Span in Isolation (ISO): We encode skill phrases t from ESCO in isolation using the aforementioned LMs, without surrounding contexts.

Average over Contexts (AOC): We leverage the surrounding context of a skill phrase t by collecting all the sentences containing t. We use all available sentences in the job postings dataset (excluding Test). For a given job posting sentence, we encode t by using one of the previous mentioned LMs. We average the embeddings of its constituent subwords to obtain the final embedding t.

Weighted Span Embedding (WSE): We obtain all inverse document frequency (idf) values of each token t_i via

$$\text{idf} = -\log \frac{n_{t_i}}{N},$$

where n_{t_i} is the number of occurrences of t_i and N the total number of tokens in our dataset. We encode the

Algorithm 1 Weakly Supervised Skill Extraction

Require: $M \in \{\text{RoBERTa}, \text{JobBERT}\}$
Require: $E \in \{\text{ISO}, \text{AOC}, \text{WSE}\}$
Require: $\tau \in [0, 1]$
  $P \leftarrow D$ ▷ A set of sentences from job postings
  $S \leftarrow S_E$ ▷ ESCO Skill embeddings of type E
  $L \leftarrow \emptyset$
  for $p \in P$ **do**
    $\theta \leftarrow 0$
    for $n \in p$ **do** ▷ Each ngram n of size $1-4$
      $E \leftarrow M(n)$
      $\Theta \leftarrow \text{CosSim}(S, E)$
      if $\max(\Theta) > \tau \wedge \max(\Theta) > \theta$ **then**
        $\theta \leftarrow \max(\Theta)$
      end if
    end for
    $L \leftarrow [L, \theta]$
  end for
  return L

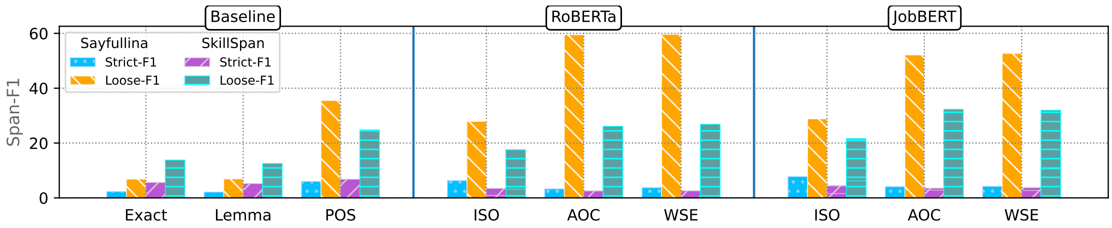

Figure 4: Results of Methods. Results on *Sayfullina* and *SkillSpan* are indicated by "Baseline" showing performance of Exact, Lemmatized (Lemma), and Part-of-Speech (POS). The performance of ISO, AOC, and WSE are separated by model, indicated by "RoBERTa" and "JobBERT". The performance of RoBERTa and JobBERT on *SkillSpan* is determined by the best performing CosSim threshold (0.8).

Table 2

Qualitative Examples of Predicted Spans. We show the gold versus predicted spans of the best performing model on both datasets. The first 5 qualitative examples are from *Sayfullina* (RoBERTa with WSE), the last 5 are from *SkillSpan*. Yellow the gold span and pink indicates the predicted span. The examples show many partial overlaps with the gold spans (but also incorrect ones), hence the high loose-F1.

	Gold	**Predicted**
Sayfullina	...a **dynamic customer focused person** to join...	...a dynamic **customer focused person to** join...
	...strong leadership and **team management skills** ...	...strong leadership **and team management skills** ...
	...speak and **written english skills** ...	...speak and **written english** skills...
	...a **team environment and working independently skills** ...	...a team **environment and working independently** skills...
	...tangible business benefit extremely **articulate** and...	...tangible **business benefit** extremely articulate and...
SkillSpan	...researcher within **machine learning** and **sensory system design** ...	...researcher within machine **learning and sensory system** design...
	...standards and procedures **accessing and updating records** ...	...standards and **procedures accessing and updating** records...
	...with a **passion for education** to...	...with a passion for **education** to...
	...understands **Agile** as a mindset...	... **understands Agile as a** mindset...
	...experience with **AWS** **GCP** **Microsoft Azure** ...	...experience **with AWS** GCP Microsoft...

input sentence and compute the weighted sum of the embeddings ($\vec{s}_j$) of the specific skill phrase in the sentence, where each t_i's IDF scores are used as weights. Again, we only use the first subword token for each tokenized word. Formally, this is

$$\vec{s}_j = \sum_{\vec{t}_i} (-\log \frac{n_{t_i}}{N}) \cdot \vec{t}_i.$$

Matching We rank pairs of ESCO embeddings $\vec{t}$ and encoded candidate n-grams $\vec{g}$ in decreasing order of cosine similarity (CosSim), calculated as

$$\text{CosSim}(\vec{t}, \vec{g}) = \frac{\vec{t}^T \vec{g}}{\|\vec{t}\| \|\vec{g}\|}.$$

We show our pseudocode of the matching algorithm in Algorithm 1. Note that in *SkillSpan* we have to set a threshold for CosSim, as there are sentences with no skills. A threshold allows us to have a "no skill" option. As seen in Figure 5, Appendix A the threshold sensitivity on *SkillSpan* differs for JobBERT: Performance fluctuates, compared to RoBERTa. Precision goes up with a higher threshold, while recall goes down. For RoBERTa, it stays similar until CosSim= 0.9. We use CosSim= 0.8 as over 2 LMs and 3 methods it provides the best cutoff.

3. Analysis of Results

Results Our main results (Figure 4) show the baselines against ISO, AOC, and WSE of both datasets. We evaluate with two types of F1, following van der Goot et al. [20]: `strict` and `loose-F1`. For full model fine-tuning, RoBERTa achieves 91.31 and 98.55 strict and loose F1 on *Sayfullina* respectively. For *SkillSpan*, this is 23.21 and 44.72 strict and loose F1 (on the available subsets of *SkillSpan*). JobBERT achieves 90.18 and 98.19 strict and loose F1 on *Sayfullina*, 49.44 and 74.41 strict and loose F1 on *SkillSpan*. The large difference between results is most likely due to lack of negatives in *Sayfullina*, i.e., all sentences contain a skill, which makes the task easier. These results highlight the difficulty of SE on *SkillSpan*, where there are negatives as well (sentences with no skills).

The exact match baseline on *SkillSpan* is higher than *Sayfullina*. We attribute this to *SkillSpan* also containing "hard skills" (e.g., "Python"), which is easier to match substrings with than "soft skills".[6]

For the performance of the skill representations on *Sayfullina*, RoBERTa and JobBERT outperform the Exact and Lemmatized baseline on strict-F1. For the POS baseline, only the ISO method of both models is slightly better. JobBERT performs better than RoBERTa in strict-F1 on both datasets.

There is a substantial difference between strict and loose-F1 on both datasets. This indicates that there is partial overlap among the predicted and gold spans. RoBERTa performs best for *Sayfullina*, achieving 59.61 loose-F1 with WSE. In addition, the best performing method for JobBERT is also WSE (52.69 loose-F1). For *SkillSpan* we see a drop, JobBERT outperforms RoBERTa with AOC (32.30 vs. 26.10 loose-F1) given a threshold of CosSim = 0.8. We hypothesize this drop in performance compared to *Sayfullina* could be attributed again to *SkillSpan* containing negative examples as well (i.e., sentences with no skill).

Qualitative Analysis A qualitative analysis (Table 2) reveals there is strong partial overlap with gold vs. predicted spans on both datasets, e.g., "...strong leadership and *team management skills*..." vs. "...strong leadership *and team management skills*...", indicating the viability of this method.

4. Conclusion

We investigate whether the ESCO skill taxonomy suits as weak supervision signal for Skill Extraction. We apply several skill representation methods based on previous work. We show that using representations of ESCO skills can aid us in this task. We achieve high loose-F1, indicating there is partial overlap between the predicted and gold spans, but need refined off-set methods to get the correct span out (e.g., human post-editing or automatic methods such as candidate filtering). Nevertheless, we see this approach as a strong alternative for supervised Skill Extraction from job postings.

Future work could include going towards multilingual Skill Extraction, as ESCO consists of 27 languages, exact matching should be trivial. For the other methods several considerations need to be taken into account, e.g., a POS-tagger and/or lemmatizer for another language and a language-specific model.

[6]The exact numbers (+precision and recall) are in Table 3, Appendix A, including the definition of strict and loose-F1.

Acknowledgments

We thank the NLPnorth group for feedback on an earlier version of this paper—in particular, Elisa Bassignana and Max Müller-Eberstein for insightful discussions. We would also like to thank the anonymous reviewers for their comments to improve this paper. Last, we also thank NVIDIA and the ITU High-performance Computing cluster for computing resources. This research is supported by the Independent Research Fund Denmark (DFF) grant 9131-00019B.

References

[1] E. Brynjolfsson, A. McAfee, Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy, Brynjolfsson and McAfee, 2011.

[2] E. Brynjolfsson, A. McAfee, The second machine age: Work, progress, and prosperity in a time of brilliant technologies, WW Norton & Company, 2014.

[3] K. Balog, Y. Fang, M. De Rijke, P. Serdyukov, L. Si, Expertise retrieval, Foundations and Trends in Information Retrieval 6 (2012) 127–256.

[4] L. Sayfullina, E. Malmi, J. Kannala, Learning representations for soft skill matching, in: International Conference on Analysis of Images, Social Networks and Texts, 2018, pp. 141–152.

[5] D. A. Tamburri, W.-J. Van Den Heuvel, M. Garriga, Dataops for societal intelligence: a data pipeline for labor market skills extraction and matching, in: 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), IEEE, 2020, pp. 391–394.

[6] M. Chernova, Occupational skills extraction with FinBERT, Master's Thesis (2020).

[7] M. Zhang, K. N. Jensen, B. Plank, Kompetencer: Fine-grained skill classification in danish job postings via distant supervision and transfer learning, Under Review, LREC 2022 (2022).

[8] M. Zhang, K. N. Jensen, S. Sonniks, B. Plank, SkillSpan: Hard and soft skill extraction from English job postings, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 4962–4984.

[9] T. Green, D. Maynard, C. Lin, Development of a benchmark corpus to support entity recognition in job descriptions, in: Proceedings of the Language Resources and Evaluation Conference, European

Language Resources Association, Marseille, France, 2022, pp. 1201–1208. URL: https://aclanthology.org/2022.lrec-1.128.

[10] A.-S. Gnehm, E. BÃ¼hlmann, S. Clematide, Evaluation of transfer learning and domain adaptation for analyzing german-speaking job advertisements, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3892–3901. URL: https://aclanthology.org/2022.lrec-1.414.

[11] A. Bhola, K. Halder, A. Prasad, M.-Y. Kan, Retrieving skills from job descriptions: A language model based extreme multi-label classification framework, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5832–5842.

[12] M. le Vrang, A. Papantoniou, E. Pauwels, P. Fannes, D. Vandensteen, J. De Smedt, Esco: Boosting job matching in europe with semantic interoperability, Computer 47 (2014) 57–64.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[14] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, B. Plank, Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 176–197.

[15] D. Zeman, J. Nivre, M. Abrams, E. Ackermann, N. Aepli, H. Aghaei, Ž. Agić, A. Ahmadi, L. Ahrenberg, C. K. Ajede, G. Aleksandravičiūtė, I. Alfina, L. Antonsen, K. Aplonova, A. Aquino, C. Aragon, M. J. Aranzabe, B. N. Arıcan, H̄. Arnardóttir, G. Arutie, J. N. Arwidarasti, M. Asahara, D. B. Aslan, L. Ateyah, F. Atmaca, M. Attia, A. Atutxa, L. Augustinus, E. Badmaeva, K. Balasubramani, M. Ballesteros, E. Banerjee, S. Bank, V. Barbu Mititelu, S. Barkarson, V. Basmov, C. Batchelor, J. Bauer, S. T. Bedir, K. Bengoetxea, G. Berk, Y. Berzak, I. A. Bhat, R. A. Bhat, E. Biagetti, E. Bick, A. Bielinskienė, K. Bjarnadóttir, R. Blokland, V. Bobicev, L. Boizou, E. Borges Völker, C. Börstell, C. Bosco, G. Bouma, S. Bowman, A. Boyd, A. Braggaar, K. Brokaitė, A. Burchardt, M. Candito, B. Caron, G. Caron, L. Cassidy, T. Cavalcanti, G. Cebiroğlu Eryiğit, F. M. Cecchini, G. G. A. Celano, S. Čéplö, N. Cesur, S. Cetin, Ö. Çetinoğlu, F. Chalub, S. Chauhan, E. Chi, T. Chika, Y. Cho, J. Choi, J. Chun, A. T. Cignarella, S. Cinková, A. Collomb, Ç. Çöltekin, M. Connor, M. Courtin, M. Cristescu, P. Daniel, E. Davidson, M.-C. de Marneffe, V. de Paiva, M. O. Derin, E. de Souza, A. Diaz de Ilarraza, C. Dickerson, A. Dinakaramani, E. Di Nuovo, B. Dione, P. Dirix, K. Dobrovoljc, T. Dozat, K. Droganova, P. Dwivedi, H. Eckhoff, S. Eiche, M. Eli, A. Elkahky, B. Ephrem, O. Erina, T. Erjavec, A. Etienne, W. Evelyn, S. Facundes, R. Farkas, M. Fernanda, H. Fernandez Alcalde, J. Foster, C. Freitas, K. Fujita, K. Gajdošová, D. Galbraith, M. Garcia, M. Gärdenfors, S. Garza, F. F. Gerardi, K. Gerdes, F. Ginter, G. Godoy, I. Goenaga, K. Gojenola, M. Gökırmak, Y. Goldberg, X. Gómez Guinovart, B. González Saavedra, B. Griciūtė, M. Grioni, L. Grobol, N. Grūzītis, B. Guillaume, C. Guillot-Barbance, T. Güngör, N. Habash, H. Hafsteinsson, J. Hajič, J. Hajič jr., M. Hämäläinen, L. Hà Mỹ, N.-R. Han, M. Y. Hanifmuti, S. Hardwick, K. Harris, D. Haug, J. Heinecke, O. Hellwig, F. Hennig, B. Hladká, J. Hlaváčová, F. Hociung, P. Hohle, E. Huber, J. Hwang, T. Ikeda, A. K. Ingason, R. Ion, E. Irimia, Ọ. Ishola, K. Ito, T. Jelínek, A. Jha, A. Johannsen, H. Jónsdóttir, F. Jørgensen, M. Juutinen, S. K, H. Kaşıkara, A. Kaasen, N. Kabaeva, S. Kahane, H. Kanayama, J. Kanerva, N. Kara, B. Katz, T. Kayadelen, J. Kenney, V. Kettnerová, J. Kirchner, E. Klementieva, A. Köhn, A. Köksal, K. Kopacewicz, T. Korkiakangas, N. Kotsyba, J. Kovalevskaitė, S. Krek, P. Krishnamurthy, O. Kuyrukçu, A. Kuzgun, S. Kwak, V. Laippala, L. Lam, L. Lambertino, T. Lando, S. D. Larasati, A. Lavrentiev, J. Lee, P. Lê Hồng, A. Lenci, S. Lertpradit, H. Leung, M. Levina, C. Y. Li, J. Li, K. Li, Y. Li, K. Lim, B. Lima Padovani, K. Lindén, N. Ljubešić, O. Loginova, A. Luthfi, M. Luukko, O. Lyashevskaya, T. Lynn, V. Macketanz, A. Makazhanov, M. Mandl, C. Manning, R. Manurung, B. Marşan, C. Mărănduc, D. Mareček, K. Marheinecke, H. Martínez Alonso, A. Martins, J. Mašek, H. Matsuda, Y. Matsumoto, A. Mazzei, R. McDonald, S. McGuinness, G. Mendonça, N. Miekka, K. Mischenkova, M. Misirpashayeva, A. Missilä, C. Mititelu, M. Mitrofan, Y. Miyao, A. Mojiri Foroushani, J. Molnár, A. Moloodi, S. Montemagni, A. More, L. Moreno Romero, G. Moretti, K. S. Mori, S. Mori, T. Morioka, S. Moro, B. Mortensen, B. Moskalevskyi, K. Muischnek, R. Munro, Y. Murawaki, K. Müürisep, P. Nainwani, M. Nakhlé, J. I. Navarro Horñiacek, A. Nedoluzhko, G. Nešpore-Bērzkalne, M. Nevaci, L. Nguyễn Thị, H. Nguyễn Thị Minh, Y. Nikaido, V. Nikolaev, R. Nitisaroj, A. Nourian, H. Nurmi, S. Ojala, A. K. Ojha, A. Olúòkun, M. Omura, E. Onwuegbuzia, P. Osenova, R. Östling, L. Øvrelid, Ş. B. Özateş, M. Özçelik, A. Özgür, B. Öztürk Başaran, H. H. Park, N. Partanen, E. Pascual, M. Passarotti, A. Patejuk, G. Paulino-Passos, A. Peljak-Łapińska, S. Peng,

C.-A. Perez, N. Perkova, G. Perrier, S. Petrov, D. Petrova, J. Phelan, J. Piitulainen, T. A. Pirinen, E. Pitler, B. Plank, T. Poibeau, L. Ponomareva, M. Popel, L. Pretkalniņa, S. Prévost, P. Prokopidis, A. Przepiórkowski, T. Puolakainen, S. Pyysalo, P. Qi, A. Rääbis, A. Rademaker, T. Rama, L. Ramasamy, C. Ramisch, F. Rashel, M. S. Rasooli, V. Ravishankar, L. Real, P. Rebeja, S. Reddy, G. Rehm, I. Riabov, M. Rießler, E. Rimkutė, L. Rinaldi, L. Rituma, L. Rocha, E. Rögnvaldsson, M. Romanenko, R. Rosa, V. Roșca, D. Rovati, O. Rudina, J. Rueter, K. Rúnarsson, S. Sadde, P. Safari, B. Sagot, A. Sahala, S. Saleh, A. Salomoni, T. Samardžić, S. Samson, M. Sanguinetti, E. Sanıyar, D. Särg, B. Saulīte, Y. Sawanakunanon, S. Saxena, K. Scannell, S. Scarlata, N. Schneider, S. Schuster, L. Schwartz, D. Seddah, W. Seeker, M. Seraji, M. Shen, A. Shimada, H. Shirasu, Y. Shishkina, M. Shohibussirri, D. Sichinava, J. Siewert, E. F. Sigurðsson, A. Silveira, N. Silveira, M. Simi, R. Simionescu, K. Simkó, M. Šimková, K. Simov, M. Skachedubova, A. Smith, I. Soares-Bastos, C. Spadine, R. Sprugnoli, S. Steingrímsson, A. Stella, M. Straka, E. Strickland, J. Strnadová, A. Suhr, Y. L. Sulestio, U. Sulubacak, S. Suzuki, Z. Szántó, D. Taji, Y. Takahashi, F. Tamburini, M. A. C. Tan, T. Tanaka, S. Tella, I. Tellier, M. Testori, G. Thomas, L. Torga, M. Toska, T. Trosterud, A. Trukhina, R. Tsarfaty, U. Türk, F. Tyers, S. Uematsu, R. Untilov, Z. Urešová, L. Uria, H. Uszkoreit, A. Utka, S. Vajjala, R. van der Goot, M. Vanhove, D. van Niekerk, G. van Noord, V. Varga, E. Villemonte de la Clergerie, V. Vincze, N. Vlasova, A. Wakasa, J. C. Wallenberg, L. Wallin, A. Walsh, J. X. Wang, J. N. Washington, M. Wendt, P. Widmer, S. Williams, M. Wirén, C. Wittern, T. Woldemariam, T.-s. Wong, A. Wróblewska, M. Yako, K. Yamashita, N. Yamazaki, C. Yan, K. Yasuoka, M. M. Yavrumyan, A. B. Yenice, O. T. Yıldız, Z. Yu, Z. Žabokrtský, S. Zahra, A. Zeldes, H. Zhu, A. Zhuravleva, R. Ziane, Universal dependencies 2.8.1, 2021. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[16] D. Kondratyuk, M. Straka, 75 languages, 1 model: Parsing universal dependencies universally, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 2779–2795.

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[19] R. Litschko, I. Vulić, S. P. Ponzetto, G. Glavaš, On cross-lingual retrieval with multilingual text encoders, Information Retrieval Journal (2022) 1–35.

[20] R. van der Goot, I. Sharaf, A. Imankulova, A. Üstün, M. Stepanovic, A. Ramponi, S. O. Khairunnisa, M. Komachi, B. Plank, From masked-language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Mexico City, Mexico, 2021.

	Dataset →	Sayfullina		SkillSpan	
	↓ **Method**, **Metric** →	**Strict** (P | R | F1)	**Loose** (P | R | F1)	**Strict** (P | R | F1)	**Loose** (P | R | F1)
Baseline	Exact	9.27 | 1.30 | 2.28	25.48 | 3.95 | 6.84	23.82 | 3.21 | 5.62	43.68 | 8.27 | 13.79
	Lemmatized	8.49 | 1.19 | 2.09	25.87 | 4.00 | 6.93	23.90 | 2.97 | 5.21	41.09 | 7.49 | 12.52
	POS	5.99 | 5.95 | 5.97	36.55 | 34.51 | 35.50	5.97 | 7.88 | 6.79	19.34 | 34.71 | 24.80
RoBERTa	ISO	6.26 | 6.25 | 6.26	26.90 | 28.98 | 27.90	2.90 | 4.24 | 3.43	12.69 | 28.61 | 17.56
	AOC	3.24 | 3.24 | 3.24	64.04 | 55.53 | 59.48	2.23 | 2.93 | 2.53	20.08 | 37.56 | 26.10
	WSE	3.67 | 3.67 | 3.67	64.64 | 55.32 | 59.61	2.29 | 2.93 | 2.57	20.90 | 37.79 | 26.85
JobBERT	ISO	7.71 | 7.72 | 7.71	27.76 | 29.95 | 28.82	4.17 | 4.65 | 4.39	17.07 | 29.48 | 21.61
	AOC	4.04 | 4.05 | 4.05	56.50 | 48.41 | 52.14	4.44 | 2.96 | 3.54	33.64 | 31.28 | 32.30
	WSE	4.15 | 4.16 | 4.15	56.98 | 49.00 | 52.69	4.78 | 3.08 | 3.74	34.01 | 30.33 | 31.95

Table 3
We show the exact numbers of the performance of the methods.

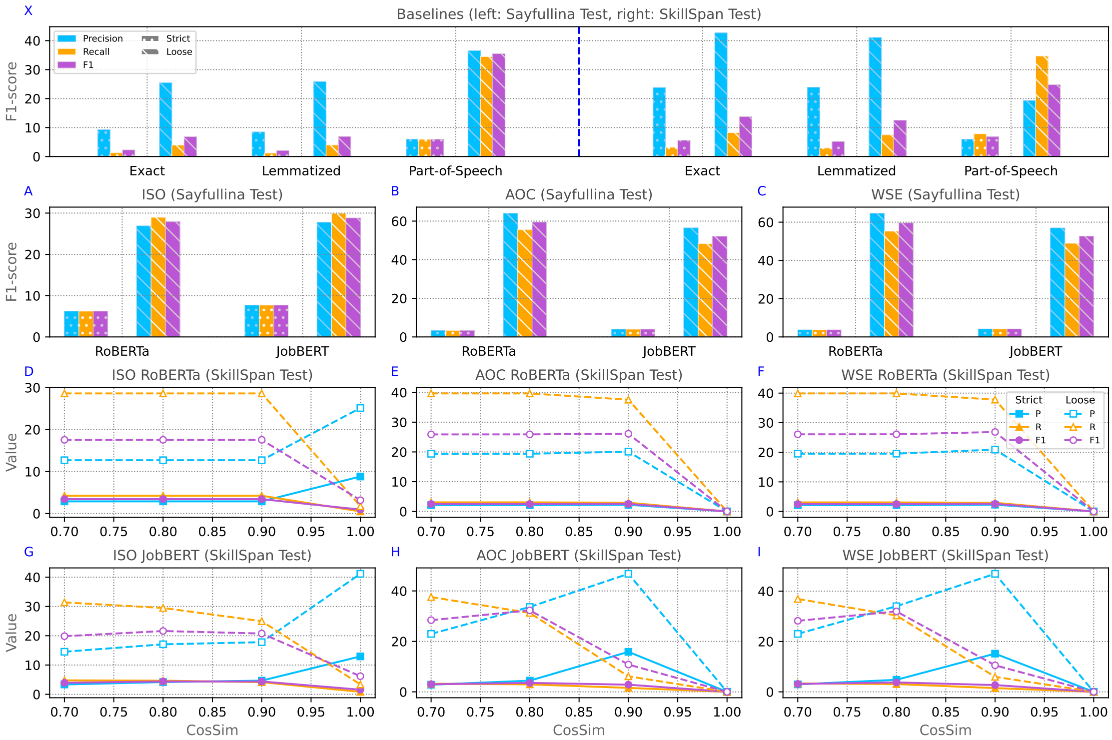

Figure 5: Results of Methods. Results of the baselines are in (X), the performance of ISO, AOC, and WSE on *Sayfullina* in (A-C), and the same performance on *SkillSpan* in (D-I) based on the model (RoBERTa or JobBERT). In D–F, we show the precision (P), recall (R), and F1 differences when taking an increasing CosSim.

A. Exact Results

Definition F1 As mentioned, we evaluate with two types of F1-scores, following van der Goot et al. [20]. The first type is the commonly used span-F1, where only the correct span and label are counted towards true positives. This is called `strict-F1`. In the second variant, we seek for partial matches, i.e., overlap between the predicted and gold span including the correct label, which counts towards true positives for precision and recall. This is called `loose-F1`. We consider the loose variant as well, because we want to analyze whether the span is "almost correct".

Exact Numbers Results We show the exact numbers of Figure 4 in Table 3 and more detailed results in Figure 5. Results show that there is high precision among the baseline approaches compared to recall. This is balanced using the representation methods for *Sayfullina*. However, we observe that there is much higher recall for *SkillSpan* than precision.