Modeling, Measuring and Exploiting Concept Drift in the Labour Market Domain

Panos Alexopoulos
Textkernel B.V.
Nieuwendammerkade 26a5
1022 AB, Amsterdam, The Netherlands
alexopoulos@textkernel.com

Spyretta Leivaditi
Kentivo B.V.
Kerksteeg 1
3582 CV, Utrecht, The Netherlands
spyretta.leivaditi@kentivo.com

ABSTRACT

The Labour Market domain is a relatively narrow domain in terms of concept types that appear in it (as it typically consists of professions, skills and qualifications) but a very broad one in terms of actual concepts (as these professions and skills can be in all kinds of domains such as Technology, Education, Finance, etc). More importantly, it is a quite volatile domain in the sense that the meaning of many concepts changes (at different rates) over time. This phenomenon, known as semantic or concept drift, poses a challenge for the maintenance and evolution of knowledge graphs that represent such domains, and requires dedicated approaches for tackling it so as to prevent such graphs from becoming irrelevant. With that in mind, in this paper we describe our experiences from dealing with concept drift in an in-house developed labour market knowledge graph, and provide insights on: i) how concept drift can be effectively defined and modeled for labour market concepts, and ii) how it can be detected, measured and effectively incorporated in the knowledge graph lifecycle.

KEYWORDS

Knowledge Graphs, Concept Drift, Labour Market

1 INTRODUCTION

A few years after Google announced that their knowledge graph allowed searching for things, not strings[1], knowledge graphs have been gaining momentum in the world's leading organisations as a means to integrate, share and exploit data and knowledge that they need in order to stay competitive [11]. Apart from Google, prominent examples of companies that develop knowledge graphs include Microsoft[2], LinkedIn[3], BBC[4] and Thomson Reuters[5]. A similar knowledge graph, for the recruitment and labour market domain, we have been developing and using for the last couple of years at Textkernel, aiming to significantly improve the way our semantic software modules parse, retrieve and match CVs and Job Vacancies.

Our knowledge graph defines and interrelates concepts and entities about the labour market and recruiting domain, such as professions, skills and qualifications, for multiple languages and countries. Using the graph, an agent (human or computer system) can answer questions like *"What are the most important skills for a certain profession?"*, *"What professions are specializations of Profession X?"* or *"What qualifications do I need in order to acquire skill Y"*. Moreover, we use the graph within our systems for a) performing entity recognition and disambiguation in CVs and vacancies and b) determining the semantic similarity between these entities when searching or matching CVs and vacancies.

Constructing the knowledge graph in an efficient and cost-effective way is a quite challenging task, not only because the labour market domain is quite broad but also because it is very heterogeneous (different industries and business areas, languages, labour markets, educational systems etc.). What is equally challenging, however, is dealing with the *concept drift* that happens to the domain's concepts as the time goes by, and causes changes to their meaning [15].

In particular, drift in our graph is mainly observed in Professions, Skills and Qualifications. Take for example journalists. Before the proliferation of the Internet and social media, a reporter would have to research stories through contacts, speaking to people, door knocking and visiting the local library to consult past publications. She would also most likely not know how to do her own video production editing but would rely on experts to do that for her. Nowadays, however, it's more likely to meet a reporter who can use effectively Google, Twitter and other modern information channels, and, to a still low yet increasing extent, data analysis and visualization tools [10]. Similar arguments can be made for other professions but also for qualifications and skills. A contemporary degree in Finance, for example, has definitely different content and even somewhat different learning objectives than it had 30 years ago. Similarly, being expert in Marketing nowadays is highly associated to being expert in Search Engine Optimization and Social Media.

These changes can be bigger or smaller, faster or slower, and more or less profound, depending on the concept type and of course the real-world dynamics. In any case, such changes can affect the quality of a knowledge graph and, therefore, dedicated frameworks for modeling, measuring and exploiting semantic drift in the context of knowledge graph maintenance and evolution are needed [14].

In this short paper, we corroborate this argument and we extend it with the following two arguments:

(1) ***The definition and modeling of semantic drift for a given knowledge graph should take into account the graph's content, domain and application context, and adapted accordingly.*** While generic formalizations of concept drift are very useful (like for example modeling drift in terms of label, intension and extension [15]), these are not necessarily directly or completely applicable to all domains and/or

[1] googleblog.blogspot.com
[2] https://arstechnica.com/information-technology/2012/06/inside-the-architecture-of-googles-knowledge-graph-and-microsofts-satori/
[3] https://engineering.linkedin.com/blog/2016/10/building-the-linkedin-knowledge-graph
[4] http://www.bbc.co.uk/ontologies
[5] https://www.scribd.com/document/288608104/Creating-the-Thomson-Reuters-knowledge-graph-and-open-permID-ODI-Summit-2015

graphs, the reason being that not all aspects of a concept's meaning contribute to its drift in the same way and to the same extent.

(2) ***There is not a unique optimal way to measure concept drift for a given knowledge graph, but rather multiple ways whose outcomes can have different interpretations and usages.*** Indeed, the values one gets when measuring concept drift can be quite different, depending on the metrics, data sources and methods/algorithms used for the measurement. Therefore, it is important that a) for a given drift measurement approach, the drift values it produces can be clearly interpreted and used, and b) for a desired interpretation/usage, an appropriate drift measurement method can be selected.

In the rest of the paper we further explain and exemplify these arguments by describing how we model and measure concept drift in our Labour Market Knowledge Graph, as well as how we apply the measurement results, not only for improving the graph but also gaining business benefits.

2 DRIFT MODELING FOR LABOUR MARKET CONCEPTS

2.1 Concept Representation

The Textkernel knowledge graph consists primarily of the following concept types:

- **Professions:** Concepts that represent groupings of jobs that involve similar tasks and require similar skills and competencies.
- **Skills:** Concepts that represent tools, techniques, methodologies, areas of knowledge, activities, and generally anything that a person can "have knowledge of", "be experienced in" or "be expert at" (e.g., Economics, Software Development, "doing sales in Africa", etc). Also concepts that represent personality traits, including communication abilities, personal habits, cognitive or emotional empathy, time management, teamwork and leadership traits (usually referred as soft skills).
- **Qualifications:** Concepts that represent *"formal outcomes of assessment and validation processes which are obtained when a competent body determines that an individual has achieved learning outcomes to given standards"* (European Qualifications Framework[6]).
- **Organizations:** Concepts that represent organizations of different types, including public organizations and institutes, private companies and enterprises, educational institutes (of all educational levels) and others.
- **Industries:** Concepts that represent industrial groupings of companies based on similar products and services, technologies and processes, markets and other criteria.

The different ways a concept can be expressed in a text (surface forms) are represented in the graph via the well-known SKOS relations *prefLabel* and *altLabel* [9]. Moreover, concepts can be taxonomically related to other concepts of the same type via the SKOS relations *broader* and *narrower* (e.g., "Software Developer" is broader than "Java Developer" and "Economics" is broader than "Microeconomics").

Additional relations are defined per concept type. In particular, professions are linked to skills and activities they involve, as well as the locations, organizations and industries where they are found. They are also linked to qualifications that are (formally or informally) required for their exercise (e.g., the BAR exam for practicing law in the United States), and, of course, to other professions that are similar to them.

Skills, in turn are linked to similar skills and activities, professions and industries they are mostly demanded by, and qualifications that develop and verify them. Finally, qualifications are linked, apart from skills, to organizations that provide them as well as the educational levels they cover.

Most of the above relations are extracted and incorporated into the knowledge graph in a semi-automatic way from a variety of structured and unstructured data sources, including CVs, Job Vacancies and Wikipedia [16] [17], as well as Search Query Logs [3]. Moreover, many of these relations are vague, i.e., there are (or could be) pairs of concepts for which it is indeterminate whether they stand in the relation or not (e.g., the similarity between different skills or the importance of a skill for a profession) [2]. The problem with vague relations is that their interpretation is highly subjective, context-dependent, and usually a matter of degree, thus making it hard to achieve a global consensus over their veracity. For this reason, in our graph, such relations have the following three properties:

- **Strength:** A number (typically from 0 to 1) indicating the strength/confidence of the relation.
- **Applicability Context:** The contexts (location, language, industry etc) in which the relation has been discovered and considered to be true.
- **Provenance:** Information about how the relation has been added to the graph (source, method, process).

These properties do not remove of course vagueness, but help towards making the relations better interpretable by both humans and systems and reducing disagreements [1]. Moreover, as we show below, these properties play an important role in the measurement of concept drift.

2.2 Concept Drift

Concept drift in the semantic knowledge representation literature is usually modeled (and measured) with respect to three aspects of a concept's meaning, namely its labels (i.e., the words used to express the concept), its intension (i.e., the concept's characteristics as expressed via its properties and relations), and its extension (i.e., the set concept's of the concept's instances) [15] [13]. The extension's role in drift is disputed by [5], suggesting that it depends on the kind of concepts under consideration.

In our knowledge graph, we adopt this latter perspective, by not considering extensions as part of our concepts' meaning and drift. One reason for that is that concepts like skills and professions are rather abstract and do not have straightforward instances (e.g., professions do not refer to specific persons or jobs). One could consider as profession instances the people that exercise them or the vacancies that are available for them, but then a change in the

[6] http://ec.europa.eu/eqf/home_en.htm

workforce size does not alter the profession's meaning. Instead, it's the qualitative characteristics of this workforce that signify a change, and that's exactly what we capture via the concepts' intension.

Nevertheless, we do not consider all properties and relations of our concepts to be part of their meaning and drift, nor to the same extent. In particular:

- We do consider as drift changes in a concept's labels, yet only when these changes are not merely additions or removals of spelling and/or morphosyntactic variations of existing labels (e.g., part-of-speech or plural form). Moreover, we consider changes in preferred labels as slightly more important than alternative labels, as the former are typically more suggestive of the concept's meaning.
- We do consider as drift changes in a concept's *broader* and *narrower* relations, with *broader* changes suggesting in general a more fundamental drift in the concept's meaning than the *narrower* ones.
- For profession concepts meaning is primarily defined by the skills and activities they involve (see the example of journalist above). Essential skills for a profession are more important than optional skills, though that can be hard to distinguish. Profession meaning also changes, though to a lesser extent, when the industries it is found in change (e.g., journalists start working in the tech sector). On the other hand, a profession concept does not drift when the locations or companies it is most popular in, change.
- For skill concepts meaning is primarily defined by their similar skills and activities, as these describe for what tasks and in what contexts a skill is used. It also changes, though to a lesser extent, when it starts being applied in different professions and industries, as part of possessing a skill includes having experience in its application contexts.
- For qualification concepts meaning is primarily defined by the skills they develop and/or verify. Secondarily, by the professions they regulate and/or are useful for (especially in some countries, qualifications are the main criterion for entering a profession).

It's worth noting that we are aware of the distinction between concept drift and concept replacement (i.e., change in the concept's core meaning) [8], but we don't really tackle this issue in our graph, because a) it can be quite difficult to define the core meaning of a concept in a way that is easily detectable, and b) it's a phenomenon that is rather rare, not causing any observable problems to our graph and its applications so far.

3 DRIFT MEASUREMENT FOR LABOUR MARKET CONCEPTS

Concept drift is typically detected and quantified by measuring the difference in meaning between two or more different versions of the same concept in different points in time [13] [12] [7] [6]. The more dissimilar the two versions are to each other, the greater the drift is.

Measuring concept meaning similarity is obviously dependent on how meaning is modeled. Thus, for example, in [15] and [13] where the authors consider as meaning the concept's intension, extension and labeling, they define corresponding similarity functions for each of these aspects. In particular, they employ string similarity metrics for measuring labeling drift, and set similarity metrics for measuring intension and extension drift. For our graph, we follow a similar approach, but with some important differences.

First, for labeling we don't use string similarity to measure change, one reason being that we don't consider spelling or morphosyntactic change as a drift. Instead, we consider labels as part of the concept's intension and we use set similarity metrics to measure the difference between a concept's changing label sets.

Second, since many of the concept relations are vague and with their validity quantified by some strength score, when we calculate similarity based on them we use metrics that can take in consideration this strength. One approach that we use, for example, is as follows: Given two versions of the same concept and a (vague) relation that influences drift, we derive the top-N related concepts for each version (based on the strength score), and we calculate their similarity using the generalized Kendall's tau [4] that can measure distance between rankings. In that way, for example, if the "Data Scientist" profession continues having the same top 10 related skills but differently ranked, a drift will be detected.

Third, in order to be able to understand and interpret concept drift better, we need a versatile measurement framework that enables the dynamic and highly configurable measurement and presentation of drift. Such a framework should take as input a set of parameters, specifying the scope, type and other characteristics of the drift we want to measure, and generate corresponding output. Examples of parameters we consider are:

- Target concept types (Professions, Skills, etc.)
- Time scope (either as a specific time period or as specific releases to be included).
- Relations and properties to be included.
- Relation applicability context and provenance.

The reason we need all these parameters, is that different values of them can yield different drift, not only in terms of intensity but also in terms of interpretation. For example, if we calculate a concept's drift using only CVs as a data source, then the drift we will measure will reflect the change in the way the workforce side of the labour market interprets and uses the concept. On the other hand, if we use only Vacancies, we shall get an idea of how the same concept changes from the industry's perspective. Similarly, if we use news articles, we will measure the change in the general perception of the concept, while the usage of more encyclopedic and definitional data sources (e.g. Wikipedia or specialized dictionaries) may indicate changes in more core aspects of the concept's meaning.

Finally, as suggested in the previous section, different relations have different influence to concept drift, and that difference needs to be considered when relation-specific drifts are aggregated. A similar argument can be made for other drift aspects like provenance or context (e.g., the change of a profession concept in a country with more advanced economy may be more important/crucial than the change in less developed country). For that reason, our drift framework supports the definition of drift aspect importance weights that are used for combining and aggregating partial drift scores.

4 DRIFT EXPLOITATION

The modeling and measurement of concept drift in our knowledge graph serves mainly two purposes, one engineering related and one of business nature. On the engineering side, the measurement and monitoring of drift helps us quantify and understand better the dynamics of our domain and our graph's content. This, in turn, enables us to plan and prioritize the maintenance and evolution of the knowledge graph much more effectively by, for example, identifying highly volatile graph aspects that need more frequent updates, and allocating more resources for that. This applies not only for computational resources (data storage capacity, data processing efficiency, etc.) but also human ones (knowledge engineers, quality analysts, annotators etc.)

On the business side, the drift in our knowledge graph indicates to a large extent the changes that take in place in the labour market, especially the one that we derive from CVs and Vacancies. These changes we can then communicate to job seekers, candidate seekers, education and training providers, policy makers, and generally anyone who can gain advantage from knowing the dynamics of the labour market.

For example, most job holders have a narrow perception of what their profession entails and to what extent and rate it evolves over time, as they usually operate in a narrow context. As a result, when these people become job seekers, they have to change this perception, otherwise they may fail to secure a new job that may have the same title but quite different content. The same applies for organizations that need to hire people but fail to do so, mainly because their job definitions are too restrictive and not in sync with the supply side of the market.

5 CONCLUSION

In this short paper we have described how we have been modeling, measuring and exploiting concept drift in a Knowledge Graph for the Labour Market domain, making the case for a more flexible, adaptable, and domain/application dependent drift tackling approach. We have shown how not all aspects of a concept's meaning contribute to its drift in the same way and to the same extent, thus requiring a careful analysis and selection of them for the domain and graph at hand. We have also shown how versatile can be the outcome of measuring concept drift (depending on the metrics, data sources and methods/algorithms used for the measurement), suggesting nevertheless that this versatility can be actually useful and, therefore, in need of proper management.

Our parameter-based drift management framework is still work in progress, requiring further research and development on how it can be properly operationalized within our enterprise. This includes full-fledged UI support, additional drift metrics, guidelines for interpreting and acting on the metrics, and a more formal user and data driven evaluation.

REFERENCES

[1] Panos Alexopoulos, Silvio Peroni, Boris Villazón-Terrazas, Jeff Z. Pan, and José Manuél Gómez-Pérez. 2014. A Metaontology for Annotating Ontology Entities with Vagueness Descriptions. In *Uncertainty Reasoning for the Semantic Web III - ISWC International Workshops, URSW 2011-2013, Revised Selected Papers.* 100–121. https://doi.org/10.1007/978-3-319-13413-0_6

[2] Panos Alexopoulos, Manolis Wallace, Konstantinos Kafentzis, and Dimitris Askounis. 2012. IKARUS-Onto: a methodology to develop fuzzy ontologies from crisp ones. *Knowl. Inf. Syst.* 32, 3 (2012), 667–695. https://doi.org/10.1007/s10115-011-0457-6

[3] Khalifeh AlJadda, Mohammed Korayem, and Trey Grainger. 2015. Improving the quality of semantic relationships extracted from massive user behavioral data. In *2015 IEEE International Conference on Big Data, Big Data 2015, Santa Clara, CA, USA, October 29 - November 1, 2015.* 2951–2953. https://doi.org/10.1109/BigData.2015.7364133

[4] Ronald Fagin, Ravi Kumar, and D. Sivakumar. 2003. Comparing Top K Lists. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '03).* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 28–36. http://dl.acm.org/citation.cfm?id=644108.644113

[5] Antske Fokkens, Serge Ter Braake, Isa Maks, and Davide Ceolin. 2016. On the Semantics of Concept Drift: Towards Formal Definitions of Concept Drift and Semantic Change. In *Proceedings of the 1st Workshop on Detection, Representation and Management of Concept Drift in Linked Open Data co-located with the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016), Bologna, Italy, November, 2016.* 10–17.

[6] Jon Atle Gulla, Geir Solskinnsbakk, Per Myrseth, Veronika Haderlein, and Olga Cerrato. 2010. Semantic Drift in Ontologies. In *WEBIST 2010, Proceedings of the 6th International Conference on Web Information Systems and Technologies, Volume 2, Valencia, Spain, April 7-10, 2010.* 13–20.

[7] Adam Jatowt and Kevin Duh. 2014. A Framework for Analyzing Semantic Change of Words Across Time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '14).* IEEE Press, Piscataway, NJ, USA, 229–238. http://dl.acm.org/citation.cfm?id=2740769.2740809

[8] Jouni-Matti Kuukanen. 2008. Makinh Sense of Conceptual Change. *History and Theory* 47, 3 (2008), 351–372.

[9] Alistair Miles and Sean Bechhofer. 2009. SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. (2009). http://www.w3.org/TR/2009/REC-skos-reference-20090818/

[10] Nic Newman. 2017. *Journalism, Media, and Technology Trends and Predictions 2017.* Technical Report. Reuters Institute for the Study of Journalism.

[11] Jeff Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu. 2017. *Exploiting Linked Data and Knowledge Graphs in Large Organisations.* Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-45654-6

[12] Gabriel Recchia, Ewan Jones, Paul Nulty, John Regan, and Peter de Bolla. 2016. Tracing Shifting Conceptual Vocabularies Through Time. In *Drift-a-LOD@EKAW (CEUR Workshop Proceedings)*, Vol. 1799. CEUR-WS.org, 2–9.

[13] Thanos G. Stavropoulos, Stelios Andreadis, Efstratios Kontopoulos, Marina Riga, Panagiotis Mitzias, and Yiannis. Kompatsiaris. SemaDrift: A Protégé Plugin for Measuring Semantic Drift in Ontologies. In *Hollink, L., Darányi, S., Meroño Peñuela, A., and Kontopoulos, E. (eds.) 1st International Workshop on Detection, Representation and Management of Concept Drift in Linked Open Data (Drift-a-LOD) in conjunction with the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW). CEUR Workshop Proceedings Vol 1799.* Bologna, Italy, 34–41.

[14] Ljiljana Stojanovic, Alexander Maedche, Boris Motik, and Nenad Stojanovic. 2002. User-Driven Ontology Evolution Management. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web (EKAW '02).* Springer-Verlag, London, UK, UK, 285–300. http://dl.acm.org/citation.cfm?id=645362.650868

[15] Shenghui Wang, Stefan Schlobach, and Michel C. A. Klein. 2011. Concept drift and how to identify it. *Journal of Web Semantics* 9 (2011), 247–265.

[16] Meng Zhao, Faizan Javed, Ferosh Jacob, and Matt McNair. 2015. SKILL: A System for Skill Identification and Normalization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15).* AAAI Press, 4012–4017. http://dl.acm.org/citation.cfm?id=2888116.2888273

[17] Wenjun Zhou, Yun Zhu, Faizan Javed, Mahmudur Rahman, Janani Balaji, and Matt McNair. 2016. Quantifying skill relevance to job titles. In *2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016.* 1532–1541. https://doi.org/10.1109/BigData.2016.7840761