Survey on the Use of Typological Information in Natural Language Processing

Helen O'Horan[1]*, **Yevgeni Berzak**[2]*, **Ivan Vulić**[1],
Roi Reichart[3], **Anna Korhonen**[1]
[1] Language Technology Lab, DTAL, University of Cambridge
[2] CSAIL MIT
[3] Faculty of Industrial Engineering and Management, Technion, IIT
helen.ohoran@gmail.com berzak@mit.edu iv250@cam.ac.uk
roiri@ie.technion.ac.il alk23@cam.ac.uk

Abstract

In recent years linguistic typology, which classifies the world's languages according to their functional and structural properties, has been widely used to support multilingual NLP. While the growing importance of typological information in supporting multilingual tasks has been recognised, no systematic survey of existing typological resources and their use in NLP has been published. This paper provides such a survey as well as discussion which we hope will both inform and inspire future work in the area.

1 Introduction

One of the biggest research challenges in NLP is the huge global and linguistic disparity in the availability of NLP technology. Still, after decades of research, high quality NLP is only available for a small number of the thousands of languages in the world. Theoretically, we have two solutions to this problem: i) development of universal, language-independent models which are equally applicable to all natural language, regardless of language-specific variation; ii) comprehensive systematisation of all possible variation in different languages.

The field of linguistic typology offers valuable resources for nearing both of these theoretical ideals: it studies and classifies world's languages according to their structural and functional features, with the aim of explaining both the common properties and the structural diversity of languages. Many of the current popular solutions to multilingual NLP: transfer of information from resource-rich to resource-poor languages (Padó and Lapata, 2005; Khapra et al., 2011; Das and Petrov, 2011; Täckström et al., 2012, inter alia), joint multilingual learning (Snyder, 2010; Cohen et al., 2011; Navigli and Ponzetto, 2012, inter alia), and development of universal models (de Marneffe et al., 2014; Nivre et al., 2016, inter alia), either assume or explicitly make use of information related to linguistic typology.

While previous work has recognised the role of linguistic typology (Bender, 2011), no systematic survey of typological information resources and their use in NLP to date has been published. Given the growing need for multilingual NLP and the increased use of typological information in recent work, such a survey would be highly valuable in guiding further development. This paper provides such a survey for *structural* typology, the areas of typological theory that consider morphosyntactic and phonological features[1], which has been the main focus of typology research in both linguistics and NLP.

We begin by introducing the field of linguistic typology and the main databases currently available (§ 2). We then discuss the role and potential of typological information in guiding multilingual NLP (§ 3). In § 4 we survey existing NLP work in terms of how typological information has been developed (4.1) and integrated in multilingual application tasks (4.2). § 5 discusses future research avenues, and § 6 concludes.

*These authors contributed equally to this work.

[1]Whilst outside of the scope of this paper, other areas of linguistic typology (e.g., lexico-semantic classifications) are also of significance for the NLP community and should be addressed in future work.

Name	Type	Coverage	Notes
World Atlas of Language Structures (WALS)	Phonology Morphosyntax Lexicosemantics	2,676 languages; 192 features; 17% of features have values	Defines language features and provides values for a large set of languages; originally intended for study of areal distribution of features
Syntactic Structures of the World's Languages (SSWL)	Morphosyntax	262 languages; 148 features; 45% of features have values	Similar to WALS, but differs in being fully open to public editing (Wikipedia-style), and by the addition of numerous example sentences for each feature
Atlas of Pidgin and Creole Language Structures (APiCS)	Phonology Morphosyntax Lexicosemantics	76 languages; 130 features; 18,526 examples	Designed to allow comparison with WALS
PHOIBLE Online	Phonology	1,672 languages; 2,160 segments	Collates and standardises several phonological segmentation databases, in addition to new data
Lyon-Albuquerque Phonological Systems Database (LAPSyD)	Phonology	422 languages	Documents a broader range of features than PHOIBLE, including syllable structures and tone systems; provides bibliographic information and links to recorded samples
URIEL Typological Compendium	Phonology Morphosyntax Lexicosemantics	8,070 languages and dialects; 284 features; approximately 439,000 feature values	Collates features from WALS, SSWL, PHOIBLE, and 'geodata'(e.g. language names, ISO codes, etc.) from sources such as Glottolog and Ethnologue; includes cross-lingual distance measures based on typological features; provides estimates for empty feature values

Table 1: An overview of major publicly accessible databases of typological information.

2 Overview of Linguistic Typology

Languages may share universal features on a deep, abstract level, but the structures found in real-world, surface-level natural language vary significantly. This variation is conventionally characterised into 'languages' (e.g. French, Hindi, Korean)[2], and linguistic typology describes how these languages resemble or differ from one another. The field comprises three pursuits: the definition of language features and their capacity for variance, the measurement and analysis of feature variance across empirical data, and the explanation of patterns observed in this data analysis. Bickel (2007) terms these three pursuits qualitative, quantitative and theoretical typology, respectively.

Typological classifications of languages have strict empirical foundations. These classifications do often support theories of causation, such as historical, areal or phylogenetic relations, but importantly, these hypotheses come second to quantitative data (Bickel, 2007). Indeed, patterns of variance may even run contrary to established theories of relations between languages based on geographical or historical proximity. For instance, Turkish and Korean are typically considered to be highly divergent in lexical features, yet their shared syntactic features make the two languages structurally quite similar. Such indications of similarity are of value for NLP which primarily seeks to model (rather than explain) cross-linguistic variation.

Typologists define and measure features according to the task at hand. Early studies, focused on word order, simply classified languages as SVO (Subject, Verb, Object), VSO, SOV, and so forth (Greenberg, 1963). There are now more various and fine-grained studies based on a wide range of features, including phonological, semantic, lexical and morphosyntactic properties (see (Bickel, 2007; Daniel, 2011) for an overview and further references). While a lot of valuable information is contained in these linguistic studies, this information is often not readily usable by NLP due to factors such as information overlap and differing definitions across studies. However, there is also a current trend towards systematically collecting typological information from individual studies in publicly-accessible databases, which are suitable for direct application in NLP (e.g., for defining features and their values).

[2]Note that there is a lacking consensus on how to define a 'language' (as opposed to a dialect, for instance) and the divisions themselves are often arbitrary and/or political. Nonetheless, the divisions are relevant insofar as they are observed in multilingual NLP.

Table 1 presents a selection of current major databases, including the Syntactic Structures of the World's Languages (SSWL) (Collins and Kayne, 2009), the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), the Phonetics Information Base and Lexicon (PHOIBLE) (Moran et al., 2014), the URIEL Typological Compendium (Littel et al., 2016), the Atlas of Pidgin and Creole Language Structures (APiCS) (Michaelis et al., 2013), and the Lyon-Albuquerque Phonological Systems Database (LAPSyD) (Maddieson et al., 2013). The table provides some basic information about these databases, including type, coverage, and additional notes. From these databases, WALS is currently by far the most commonly-used typological resource in NLP due to its broad coverage of features and languages.

We next discuss the potential of typological information to guide multilingual NLP and the means by which this can be done.

3 Multilingual NLP and the Role of Typologies

The recent explosion of language diversity in electronic texts has made it possible for NLP to move increasingly towards multilingualism. The biggest challenge in this pursuit has been resource scarcity. In order to achieve high quality performance, NLP algorithms have relied heavily on manually crafted resources such as large linguistically-annotated datasets (treebanks, parallel corpora, etc.) and rich lexical databases (terminologies, dictionaries, etc.). While such resources are available for key NLP tasks (POS tagging, parsing, etc.) in well-researched languages (e.g. English, German, and Chinese), for the majority of other languages they are lacking altogether. Since resource creation is expensive and cannot be realistically carried out for all tasks in all languages, much recent research in multilingual NLP has investigated ways of overcoming the resource problem.

One avenue of research that aims to solve this problem has been unsupervised learning, which exploits unlabelled data that is now available in multiple languages. Over the past two decades increasingly sophisticated unsupervised methods have been developed and applied to a variety of tasks and in some cases also to multiple languages (Cohen and Smith, 2009; Reichart and Rappoport, 2010; Snyder, 2010; Spitkovsky et al., 2011; Goldwasser et al., 2011; Baker et al., 2014, inter alia). However, while purely unsupervised approaches are appealing in side-stepping the resource problem, their relatively low performance has limited their practical usefulness (Täckström et al., 2013). More success has been gained with solutions that use some form of supervision or guidance to enable NLP for less-resourced languages (Naseem et al., 2010; Zhang et al., 2012; Täckström et al., 2013, inter alia). In what follows, we consider three such solutions: language transfer, joint multilingual learning, and the development of universal models. We discuss the guidance employed in each, paying particular attention to typological guidance.

Language Transfer This very common approach exploits the fact that rich linguistic resources do exist for some languages. The idea is to use them for less-resourced languages via data (i.e. parallel corpora) and/or model transfer. This approach has been explored widely in NLP (Hwa et al., 2005; McDonald et al., 2011; Petrov et al., 2012; Zhang and Barzilay, 2015). It has been particularly popular in recent research on dependency parsing, where a variety of methods have been explored. For example, most work for resource-poor languages has combined delexicalised parsing with cross-lingual transfer (e.g. (Zeman and Resnik, 2008; McDonald et al., 2011; Søgaard, 2011; Rosa and Zabokrtsky, 2015)). Here, a delexicalised parser is first trained on a resource-rich source language, with both languages POS-tagged using the same tagset, and then applied directly to a resource-poor target language.

While such a transfer approach outperforms unsupervised learning, it does not achieve optimal performance. One potential reason for this is that the tagset used by a POS tagger may not fit a target language which exhibits significantly different morphological features to a source language for which the tagset was initially developed (Petrov et al., 2012). Although parallel data can be used to give additional guidance which improves transfer (McDonald et al., 2011), such data are only available for some language pairs and cannot be used in truly resource-poor situations.

An alternative direction that has recently emerged uses typological information as a form of non-parallel guidance in transfer. This direction capitalises on the fact that languages do exhibit systematic cross-lingual connections at various levels of linguistic description (e.g. similarities in language structure), despite their great diversity. Captured in typological classifications at the level of generalisation useful

for NLP, such information can be used to support multilingual NLP in a variety of ways (Bender, 2011). For example, it can be used to define the similarity between two languages with respect to the linguistic information one hopes to transfer; it can also help to define the optimal degree, level and method of transfer. For example, direct transfer of POS tagging is more likely to succeed when languages are sufficiently similar in terms of morphology in particular (Hana et al., 2004; Wisniewski et al., 2014).

Typological information has been used to guide language transfer mostly in the areas of part-of-speech tagging and parsing, e.g. (Cohen and Smith, 2009; McDonald et al., 2011; Berg-Kirkpatrick and Klein, 2010; Naseem et al., 2012; Täckström et al., 2013). Section 4 surveys such works in more detail.

Multilingual Joint Learning Another approach involves learning information for multiple languages simultaneously, with the idea that the languages will be able to support each other (Snyder, 2010; Navigli and Ponzetto, 2012). This can help in the challenging but common scenario where none of the languages involved has adequate resources. This applies even with English, where annotations needed for training basic tools are primarily available only for newspaper texts and a handful of other domains. In some areas of NLP, e.g. word sense disambiguation (Navigli and Ponzetto, 2012), multilingual learning has outperformed independent learning even for resource-rich languages, with larger gains achieved by increasing the number of languages.

Success has also been achieved on morphosyntactic tasks. For example, Snyder (2010) observes that cross-lingual variations in linguistic structure correspond to systematic variations in ambiguity, so that what one language leaves implicit, another one will not. For instance, a given word may be tagged as either a verb or a noun, yet its equivalent in other languages may not present such ambiguity. Together with his colleagues, Snyder exploited this variation to improve morphological segmentation, POS tagging, and syntactic parsing for multiple languages. Naseem et al. (2012) introduced a selective sharing approach to improve multilingual dependency parsing where the model first chooses syntactic dependents from all the training languages and then selects their language-specific ordering to tie model parameters across related languages. Because the ordering decisions are influenced by languages with similar properties, this cross-lingual sharing is modelled using typological features. In such works, typological information has been used to facilitate the matching of structural features across languages, as well as in the selection of languages between which linguistic information should be shared.

Development of Universal Models A long-standing goal that has gained renewed interest recently is the development of language-independent (i.e. *universal*) models for NLP (Bender, 2011; Petrov et al., 2012). Much of the recent interest has been driven by the Universal Dependencies (UD) initiative. It aims to develop cross-linguistically consistent treebank annotation for many languages for the purposes of facilitating multilingual parser development and cross-lingual learning (Nivre et al., 2016). The annotation scheme is largely based on universal Stanford dependencies (de Marneffe et al., 2014) and universal POS tags (Petrov et al., 2012). UD treebanks have been developed for 40 languages to date. Whilst still biased towards contemporary Indo-European languages, the collection developed by this initiative is gradually expanding to include additional language families.

The development of a truly universal resource will require taking into account typological variation for optimal coverage. For example, while the current UD scheme allows for language-specific tailoring, in the future, language type-specific tailoring may offer a useful alternative, aligned with the idea of universal modeling (Bender, 2011).

4 Development and Uses of Typological Information in NLP

Given the outlined landscape of multilingual NLP and its relation to structural typology, we now survey existing approaches for obtaining (4.1) and utilizing (4.2) typological information to support various NLP tasks.

4.1 Development of Typological Information for NLP

Typological information has been obtained using two main approaches: i) extraction from manually constructed linguistic resources, such as the databases reviewed in §2; and ii) automatic learning. The two

methods have been used independently and in combination, and both are based on the assumption (be it explicit or implicit) that typological relations may be fruitfully used in NLP.

Manual Extraction from Linguistic Resources Manually crafted linguistic resources – in particular the WALS database – have been the most commonly used sources of typological information in NLP. To date, syntactic parsing (Naseem et al., 2012; Täckström et al., 2013; Zhang and Barzilay, 2015; Ammar et al., 2016) and POS tagging (Zhang et al., 2012; Zhang et al., 2016) were the predominant areas for integration of structural information from such databases. In the context of these tasks, the most frequently used features related to word ordering according to coarse syntactic categories. Additional areas with emerging research which leverages externally-extracted typological features are phonological modeling (Tsvetkov et al., 2016; Deri and Knight, 2016) and language learning (Berzak et al., 2015).

While information obtained from typological databases has been successfully integrated in several NLP tasks, a number of challenges remain. Perhaps the most crucial challenge is the partial nature of the documentation available in manually-constructed resources. For example, WALS currently covers about 17% of its possible feature values (Dryer and Haspelmath, 2013) (see Table 1 for feature coverage of other typological databases). The integration of information from different databases is challenging due to differences in feature taxonomies as well as information overlap across repositories. Furthermore, available typological classifications contain different feature types, including nominal, ordinal and interval variables, and features that mix several types of values. This property hinders systematic and efficient encoding of such features in NLP models – a problem which thus far has only received a partial solution in the form of feature binarisation (Georgi et al., 2010). Further, typological databases are constructed manually using limited resources, and do not contain information on the distribution of feature values within a given language. This results in incomplete feature characterisations, as well as inaccurate generalisations. For example, WALS encodes only the dominant noun-adjective ordering for French, although in some cases this language also permits the adjective-noun ordering.

Other aspects of typological databases may require feature pruning and preprocessing prior to use. For example, some features in WALS, such as feature 81B "Languages with two Dominant Orders of Subject, Object, and Verb" are applicable only to a subset of the world's languages. Currently, no explicit specification for feature applicability is present in WALS or other typological resources. Furthermore, distinct features may encode overlapping information, as in the case of WALS features 81A "Order of Subject Verb and Object" and 83A "Order of Verb and Object", where the latter can be deduced from the former. Although many of these issues have been noted in previous research (Östling, 2015), there are currently no standard procedures for preprocessing typological databases for NLP use.

Despite the caveats presented above, typological resources do offer an abundance of valuable structural information which can be integrated in many NLP tasks. This information is currently substantially underutilised. Out of 192 available features in WALS, only a handful of word order features are typically used to enhance multilingual NLP. Meanwhile, the complementary information on additional languages and feature types offered by other repositories has, to our knowledge, rarely been exploited in NLP. This readily-available information could be used more extensively in tasks such as POS tagging and syntactic parsing, which have already gained from typological knowledge, and it could also be used to support additional areas of NLP.

Automatic Learning of Typological Information The partial coverage of existing typological resources, stemming from the difficulty of obtaining such information manually, have sparked a line of work on automatic acquisition of typological information. Here too, WALS has been the most common reference for defining the features to be learned.

Several approaches were introduced for automatic induction of typological information through multilingual word alignments in parallel texts. Mayer and Cysouw (2012) use alignments to induce language similarities, and use this approach to support learning of fine-grained features, such as the typology of person interrogatives (e.g., English "who"). In Östling (2015) multilingual word alignments are used to project POS tags and syntactic trees for translations of the New Testament, and subsequently learn typological information relating to word order. The predicted typological features, when evaluated against

WALS, achieve high accuracy. This method not only extends WALS word order documentation to hundreds of new languages, but also quantifies the frequency of different word orders across languages – information that is not available in manually crafted typological repositories.

Typological information can also be extracted from Interlinear Glossed Text (IGT). Such resources contain morphological segmentation, glosses and English translations of example sentences collected by field linguists. Lewis and Xia (2008) and Bender et al. (2013) demonstrate that IGT can be used to extract typological information relating to word order, case systems and determiners for a variety of languages.

Another line of work seeks to increase the coverage of typological information using existing information in typological databases. Daumé III and Campbell (2007) and Bakker (2008) use existing WALS features to learn typological implications of the kind pioneered by Greenberg (1963). Such rules can then be used to predict unknown feature values for new languages. Georgi et al. (2010) use documented WALS features to cluster languages, and subsequently predict new feature values using nearest-neighbour projection. A classifier-based approach for predicting new feature values from documented WALS information is presented in (Takamura et al., 2016). Coke et al. (2016) predict word order typological features by combining documented typological and genealogical features with the multilingual alignment approach discussed above.

An alternative approach for learning typological information uses English as a Second Language (ESL) texts (Berzak et al., 2014). This work demonstrates that morphosyntactic typological similarities between languages are largely preserved in second language structural usage. It leverages this observation to approximate typological similarities between languages directly from ESL usage patterns and further utilise these similarities for nearest neighbor prediction of typological features. The method evaluates competitively compared to baselines in the spirit of (Georgi et al., 2010) which rely on existing typological documentation of the target language for determining its nearest neighbors.

In addition, a number of studies learned typological information tailored to the particular task and data at hand (i.e. *task-based development*). For example, Song and Xia (2014) process Ancient Chinese using Modern Chinese parsing resources. They manually identify and address statistical patterns in variation between monolingual corpora in each language, and ultimately optimise the model performance by selectively using only the Modern Chinese features which correspond to Ancient Chinese features.

Although automatically-learned typological classifications have not been used frequently to date, they hold great promise for extending the use of typological information in NLP. Furthermore, such work offers an additional axis of interaction between linguistic typology and NLP, namely using computational modeling in general and NLP in particular to assist linguistic documentation and analysis of typological information. We discuss the future prospects of these research directions in § 6.

4.2 Uses of Typological Information in NLP

Multilingual Syntactic Parsing As mentioned in § 4.1, the main area of NLP in which information from structural typology has been exploited thus far is multilingual dependency parsing. In this task, a priori information about the predominant orderings of syntactic categories across languages are used to guide models when parsing a resource-poor language and using training data from other languages. This information is available in typological resources (e.g., WALS) which, among a variety of other syntactic features, list the dominant word orderings for many languages (see Table 1).

A seminal work that integrates typological word order information in multilingual dependency parsing (Naseem et al., 2012) presents the idea of "selective sharing" between source and target languages. In brief, while the identity of possible dependents for a given syntactic category is (hypothesised to be) language-universal, their ordering is language-specific. The work then presents a generative multilingual parsing model in which dependent ordering parameters are conditioned on word order typology, obtained from WALS. Specifically, the paper utilises the following word order features (henceforth WALS Basic word Order, WBO): 81A (Subject Verb and Object), 85A (Adposition and Noun), 86A (Genitive and Noun), 87A (Adjective and Noun), 88A (Demonstrative and Noun) and 89A (Numeral and Noun). This information enables the model to take into account dependent orderings only when the source language has a similar word order typology to the target language. In a similar vain, Täckström et al. (2013) present

an instance of the typologically guided selective sharing idea within a discriminative parsing framework. They group the model features into features that encode arc directionality and word order, and those that do not. The former group is then coupled with the same WBO features used by Naseem et al. (2012) via feature templates that match the WALS properties with their corresponding POS tags. Additional features that group languages according to combinations of WALS features as well as coarse language groups (Indo-European versus Altaic), result in further improvements in parsing performance.

Zhang and Barzilay (2015) extended the selective sharing approach for discriminative parsing to tensor-based models using the same WBO features as in (Naseem et al., 2012) and (Täckström et al., 2013). While traditional tensor-based parsers represent and assign non-zero weights to all possible combinations of atomic features, this work presents a hierarchical architecture that enables discarding chosen feature combinations. This allows the model to integrate prior typological knowledge, while ignoring uninformative combinations of typological and dependency features. At the same time, it capitalises on the automatisation of feature construction inherent to tensor models to generate combinations of informative typology-based features, further enhancing the added value of typological priors.

Another successful integration of externally-defined typological information in parsing is the work of Ammar et al. (2016). They present a multilingual parser trained on a concatenation of syntactic treebanks of multiple languages. To reduce the adverse impact of contradicting syntactic information in treebanks of typologically distinct languages, while still maintaining the benefits of additional training data for cross-linguistically consistent syntactic patterns, the parser encodes a language-specific bias for each given input language. This bias is based on the identity of the language and its WBO features as used in (Naseem et al., 2012; Täckström et al., 2013; Zhang and Barzilay, 2015). Differently from prior work, their parsing model also encodes all other features in the WALS profile of the relevant language. Overall, this strategy leads to improved parsing performance compared to monolingually trained baseline parsers.

While the papers surveyed above use prior information about word order typology extracted from WALS, word order information for guiding multilingual parsing can also be extracted in a bottom-up, data-driven fashion, without explicit reference to typological taxonomies. For example, in Søgaard (2011), training sentences in a source language are selected based on the perplexity of their coarse POS tag sequence under a target language POS language model. This approach essentially chooses sentences that exhibit similar word orderings in both source and target languages, thus realizing a bottom-up variant of the typology-based selective sharing methods discussed above.

There are also several methods which have made use of less explicit typological information. For instance, Berg-Kirkpatrick and Klein (2010) selectively combine languages in their method for cross-lingual dependency grammar induction using a phylogeny tree, which has been constructed from external (unspecified) knowledge of language families. Zeman and Resnik (2008) demonstrate improved performance of cross-lingually transferred dependency parsers within sets of typologically similar languages (e.g. Swedish-Danish, Hindi-Urdu); they do not explain how languages may be determined as "closely-related", though presumably this decision was based on the intuition of the researchers or on widely-acknowledged generalisations.

POS Tagging, Phonological Modeling and Language Learning Besides dependency parsing, several other areas have started integrating typological information in various forms. A number of such works revolve around the task of POS tagging. For example, in Zhang et al. (2012), the previously discussed WBO features were used to inform mappings from language-specific to a universal POS tagset. In (Zhang et al., 2016), WBO feature values are used to evaluate the quality of a multilingual POS tagger.

Another application area which benefited from integration of typological knowledge are phonological models of text. In (Tsvetkov et al., 2016) a multilingual neural phoneme-based language model is trained on several languages using a shared phonological inventory. The model is conditioned on the identity of the language at hand, as well as its phonological features obtained from a concatenation of phonological features from WALS, PHOIBLE and Ethnologue, extracted from URIEL. The resulting model subsumes and outperforms monolingually trained models for phone sequence prediction. Deri and Knight (2016) use URIEL to obtain phone and language similarity metrics, which are used for adjusting Grapheme to Phoneme (G2P) models from resource rich to resource poor languages.

Berzak et al. (2015) use typological classifications to study language learning. Formalizing the theory of "Contrastive Analysis" which aims to analyse learning difficulties in a foreign language by comparing native and foreign language structures, they build a regression model that predicts language-specific grammatical error distributions by comparing typological features in the native and foreign languages.

5 Typological Information and NLP: What's Next?

§ 4.2 surveyed the current uses of typological information in NLP. Here we discuss several future research avenues that might benefit from tighter integration of linguistic typologies and multilingual NLP.

Encoding Typological Information in Traditional Machine Learning-based NLP One of the major open challenges for typologically-driven NLP is the construction of principled mechanisms for the integration of typological knowledge in machine learning-based algorithms. Here, we briefly discuss a few traditional machine learning frameworks which support encoding of expert information, and as such hold promise for integrating typological information in NLP.

Encoding typological knowledge into machine learning requires mechanisms that can bias *learning (parameter estimation)* and *inference (prediction)* of the model towards predefined knowledge. Algorithms such as the structured perceptron (Collins, 2002) and structured SVM (Taskar et al., 2004) iterate between an inference step and a parameter update step with respect to gold training labels. The inference step is a natural place for encoding external knowledge through constraints. It biases the prediction of the model to agree with external knowledge, which, in turn, affects both the training process and the final model prediction. As typological information often reflects tendencies rather than strict rules, *soft constraints* are helpful. Ultimately, an efficient mechanism for encoding soft constraints into the inference step is needed. Indeed, several modeling approaches have been proposed that do exactly this: constraint-driven learning (CODL) (Chang et al., 2007), posterior regularisation (PR) (Ganchev et al., 2010), generalized expectation (GE) (Mann and McCallum, 2008), and dual decomposition (Globerson and Jaakkola, 2007), among others. Such approaches have been applied successfully to various NLP tasks where external knowledge is available. Examples include POS tagging and parsing (Rush et al., 2010; Rush et al., 2012), information extraction (Riedel and McCallum, 2011; Reichart and Barzilay, 2012), and discourse analysis (Guo et al., 2013), among others. In addition to further extensions to the modeling approaches surveyed in §4.2, these type of frameworks could expedite principled integration of typological information in NLP.

Typologies and Multilingual Representation Learning While the traditional models surveyed above assume a predefined feature representation and focus on generating the best prediction of the output labels, a large body of recent NLP research has focused on learning dense real-valued vector representations —- i.e., word embeddings (WEs). WEs serve as pivotal features in a range of downstream NLP tasks such as parsing, named entity recognition, and POS tagging (Turian et al., 2010; Collobert et al., 2011; Chen and Manning, 2014). The extensions of WE models in bilingual and multilingual settings (Klementiev et al., 2012; Hermann and Blunsom, 2014; Coulmance et al., 2015; Vulić and Moens, 2016, inter alia) abstract over language-specific features and attempt to represent words from several languages in a language-agnostic manner such that similar words (regardless of the actual language) obtain similar representations. Such multilingual WEs facilitate cross-lingual learning, information retrieval and knowledge transfer. The extent to which multilingual WEs capture word meaning across languages has been recently evaluated in (Leviant and Reichart, 2015) with the conclusion that multilingual training usually improves the alignment between the induced WEs and the meaning of the participating words in each of the involved languages.

Naturally, as these models become more established and better understood, the challenge of external knowledge encoding becomes more prominent. Recent work has examined the ability to map from word embeddings to interpretable typological representations (Qian et al., 2016). Furthermore, a number of works (Faruqui et al., 2015; Rothe and Schütze, 2015; Osborne et al., 2016; Mrkšić et al., 2016) proposed means through which external knowledge from structured knowledge bases and specialised linguistic resources can be encoded in these models. The success of these works suggests that more extensive integration of external linguistic knowledge in general, and typological knowledge in particular, is likely to play a key role in the future development of WE representations.

Can NLP Support Typology Construction? As discussed in §4, typological resources are commonly constructed manually by linguists. Despite the progress made in recent years in the digitisation and collection of typological knowledge in centralised repositories, their coverage remains limited. Following the work surveyed in §4.1 on automatic learning of typological information, we believe that NLP could play a much larger role in the study of linguistic typology and the expansion of such resources. Future work in these directions will not only assist in the global efforts for language documentation, but also substentially extend the usability of such resources for NLP purposes.

6 Commentary; conclusion

This paper has provided a survey of linguistic typologies and the many recent works in multilingual NLP that have benefited from such resources. We have shown how combined knowledge of linguistic universals and typological variation has been used to improve NLP by enabling the use of cross-linguistic data in the development and application of resources. Promising examples of both explicit and implicit typological awareness in NLP have been presented. We have concluded with a discussion on how typological information could be used to inform improved experimental and conceptual practice in NLP. We hope that this survey will be useful in both informing and inspiring future work on linguistic typologies and multilingual NLP.

Acknowledgments

This work is supported by ERC Consolidator Grant LEXICAL (no 648909) and by the Center for Brains, Minds and Machines (CBMM) funded by the NSF STC award CCF-1231216. The authors are grateful to the anonymous reviewers for their helpful comments and suggestions.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *TACL*.

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *EMNLP*, pages 278–289.

Dik Bakker. 2008. INFER: Inferring implications from the WALS database. *Sprachtypologie und Universalienforschung*, 3(61):186–198.

Emily M. Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *LaTeCH 2013*.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 3(6):1–26.

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *ACL*, pages 1288–1297.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. In *CoNLL*, pages 21–29.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2015. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *CoNLL*, pages 94–102.

Balthasar Bickel. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology*, 11(1):239–251.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *ACL*, pages 280–287.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

Shay Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *NAACL-HLT*, pages 74–82.

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *EMNLP*, pages 50–61.

Reed Coke, Ben King, and Dragomir R. Radev. 2016. Classifying syntactic regularities for hundreds of languages. *CoRR*, abs/1603.08016.

Chris Collins and Richard Kayne. 2009. Syntactic structures of the world's languages. `http://sswl.railsplayground.net/`.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *EMNLP*, pages 1–8.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word embeddings. In *EMNLP*, pages 1109–1113.

Michael Daniel. 2011. Linguistic typology and the study of language. In *The Oxford Handbook of Linguistic Typology*, pages 43–68.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*, pages 600–609.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *ACL*, pages 65–72.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *LREC*, pages 4585–4592.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *ACL*, pages 399–408.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *NAACL-HLT*, pages 1606–1615.

Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *COLING*, pages 385–393.

Amir Globerson and Tommi S Jaakkola. 2007. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *NIPS*, pages 553–560.

Dan Goldwasser, Roi Reichart, James Clarke, and Dan Roth. 2011. Confidence driven unsupervised semantic parsing. In *ACL*, pages 1486–1495.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of Language*, 2:73–113.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2013. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *NAACL-HLT*, pages 928–937.

Jiri Hana, Anna Feldman, and Chris Brew. 2004. A resource-light approach to Russian morphology: Tagging Russian using Czech resources. In *EMNLP*, pages 222–229.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *ACL*, pages 58–68.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara I. Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Mitesh M. Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011. Together we can: Bilingual bootstrapping for WSD. In *ACL*, pages 561–569.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

William D Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *IJCNLP*, pages 685–690.

Patrick Littel, David R. Mortensen, and Lori Levin. 2016. URIEL Typological database. Pittsburgh: CMU.

Ian Maddieson, Sébastien Flavier, Egidio Marsico, Christophe Coupé, and François Pellegrino. 2013. LAPSyd: Lyon-Albuquerque phonological systems database. In *INTERSPEECH*, pages 3022–3026.

Gideon S. Mann and Andrew McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *ACL*, pages 870–878.

Thomas Mayer and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 54–62.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*, pages 62–72.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *Atlas of Pidgin and Creole Language Structures Online*.

Steven Moran, Daniel McCloy, and Richard Wright, editors. 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *NAACL-HLT*, pages 142–148.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proc. of EMNLP 2010*.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *ACL*, pages 629–637.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Joining forces pays off: Multilingual joint word sense disambiguation. In *EMNLP-CoNLL*, pages 1399–1410.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*, pages 1659–1666.

D. Osborne, S. Narayan, and S. B. Cohen. 2016. Encoding prior knowledge with eigenword embeddings. *Transactions of the ACL (to appear)*.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *ACL*, pages 205–211.

Sebastian Padó and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *EMNLP*, pages 859–866.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *LREC*, pages 2089–2096.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Investigating language universal and specific properties in word embeddings. In *ACL*, pages 1478–1488.

Roi Reichart and Regina Barzilay. 2012. Multi event extraction guided by global constraints. In *NAACL*, pages 70–79.

Roi Reichart and Ari Rappoport. 2010. Improved fully unsupervised parsing with zoomed learning. In *EMNLP*, pages 684–693.

Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *EMNLP*, pages 1–12.

Rudolf Rosa and Zdenek Zabokrtsky. 2015. KLcpos3 - a language similarity measure for delexicalized parser transfer. In *ACL*, pages 243–249.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *ACL*, pages 1793–1803.

Alexander M Rush, David Sontag, Michael Collins, and Tommi Jaakkola. 2010. On dual decomposition and linear programming relaxations for natural language processing. In *EMNLP*, pages 1–11.

Alexander M. Rush, Roi Reichart, Michael Collins, and Amir Globerson. 2012. Improved parsing and POS tagging using inter-sentence consistency constraints. In *EMNLP-CoNLL*, pages 1434–1444.

Ben Snyder. 2010. *Unsupervised Multilingual Learning*. PhD thesis. MIT.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *ACL*, pages 682–686.

Yan Song and Fei Xia. 2014. Modern Chinese helps archaic Chinese processing: Finding and exploiting the shared properties. In *LREC*, pages 3129–3136.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2011. Lateen EM: Unsupervised training with multiple objectives, applied to dependency grammar induction. In *EMNLP*, pages 1269–1280.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL-HLT*, pages 477–487.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *NAACL-HLT*, pages 1061–1071.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *LREC*, pages 69–76.

Ben Taskar, Carlos Guestrin, and Daphne Koller. 2004. Max-margin Markov networks. In *NIPS*, pages 25–32.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W. Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *NAACL*, pages 1357–1366.

Joseph P. Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL*, pages 384–394.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *EMNLP*, pages 1779–1785.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pages 35–42.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *EMNLP*, pages 1857–1867.

Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to map into a universal POS tagset. In *EMNLP*, pages 1368–1378.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *NAACL*, pages 1307–1317.