Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology

Emily M. Bender
University of Washington
Seattle, WA, USA
ebender@u.washington.edu

Abstract

In this position paper, I argue that in order to create truly language-independent NLP systems, we need to incorporate linguistic knowledge. The linguistic knowledge in question is not intricate rule systems, but generalizations from linguistic typology about the range of variation in linguistic structures across languages.

1 Introduction

Language independence is commonly presented as one of the advantages of modern, machine-learning approaches to NLP. Once an algorithm is developed, the argument goes, it can trivially be extended to another language; "all" that is needed is a suitably large amount of training data for the new language.[1] This is indeed a virtue. However, the typical approach to developing language-independent systems is to eschew using any linguistic knowledge in their production. In this position paper, I argue that, on the contrary, the production of language-independent NLP technology *requires* linguistic knowledge, and that the relevant kind of linguistic knowledge is in fact relatively inexpensive.

The rest of this paper is structured as follows: In Section 2, I discuss how linguistically naïve systems can end up tuned to the languages they were originally developed for. In Section 3, I survey the long papers from ACL2008:HLT to give a snapshot of how linguistic diversity is currently handled in our field. In Section 4, I give a brief overview of Linguistic Typology, and suggest how knowledge derived from this field can be profitably incorporated into language-independent NLP systems.

2 Hidden Language Dependence

A simple example of subtle language dependence is the way in which n-gram models work better for languages that share important typological properties with English. On the face of it, n-gram models code in no linguistic knowledge. They treat natural language text as simple sequences of symbols and automatically reflect the "hidden" structure through the way it affects the distributions of words in various (flat, unstructured) contexts. However, the effectiveness of n-gram models in English (and similar languages) is partially predicated on two properties of those languages: relatively low levels of inflectional morphology, and relatively fixed word order.

As is well-known by now, languages with more elaborate morphology (more morphemes per word, more distinctions within the same number of morphological slots, and/or fewer uninflected words) present greater data sparsity problems for language models. This data sparsity limits the ability of n-gram models to capture the dependencies between open-class morphemes, but also closed class morphemes. The information expressed by short function words in English is typically expressed by the inflectional morphology in languages with more elaborate morphological systems. Word-based n-gram models have no way of representing the function morphemes in such a language. In addition, for n-gram models to capture inter-word dependencies, both words have to appear in the n-gram window. This will happen more consistently in languages with relatively fixed word order, as compared to languages with relatively free word order.

Thus even though n-grams models can be built

[1] This of course abstracts away from the production of such data, which may require both significant pre-processing and annotation work. For the purposes of the present argument, however, we can assume that all language-independent NLP systems are unicode-enabled, assume a definition of "word" that is cross-linguistically applicable, and require the type of annotations that are likely to have already been deployed for another purpose.

Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics, pages 26–32,
Athens, Greece, 30 March, 2009.

without any hand-coding of linguistic knowledge, they are not truly language independent. Rather, their success depends on typological properties of the languages they were first developed for. A more linguistically-informed (and thus more language independent) approach to n-gram models is the factored language model approach of Bilmes and Kirchhoff (2003). Factored language models address the problems of data-sparsity in morphologically complex languages by representing words as bundles of features, thus capturing dependencies between subword parts of adjacent words.

A second example of subtle language dependence comes from Dasgupta and Ng (2007), who present an unsupervised morphological segmentation algorithm meant to be language-independent. Indeed, this work goes much further towards language independence than is the norm (see Section 3). It is tested against data from English, Bengali, Finnish and Turkish, a particularly good selection of languages in that it includes diversity along a key dimension (degree of morphological complexity), as well as representatives of three language families (Indo-European, Uralic, and Altaic). Furthermore, the algorithm is designed to detect more than one prefix or suffix per word, which is important for analyzing morphologically complex languages. However, it seems unrealistic to expect a one-size-fits-all approach to be achieve uniformly high performance across varied languages, and, in fact, it doesn't. Though the system presented in (Dasgupta and Ng, 2007) outperforms the best systems in the 2006 PASCAL challenge for Turkish and Finnish, it still does significantly worse on these languages than English (F-scores of 66.2 and 66.5, compared to 79.4).

This seems to be due to an interesting interaction of at least two properties of the languages in question. First, the initial algorithm for discovering candidate roots and affixes relies on the presence of bare, uninflected roots in the training vocabulary, extracting a string as a candidate affix (or sequence of affixes) when it appears at the end (or beginning) of another string that also appears independently. In Turkish and Finnish, verbs appear as bare roots in many fewer contexts than in English.[2] This is also true in Bengali, and the authors note that their technique for detecting allomorphs is critical to finding "out-of-vocabulary" roots (those unattested as stand-alone words) in that language. However, the technique for finding allomorphs assumes that "roots exhibit the character changes during attachment, not suffixes" (p.160), and this is where another property of Finnish and Turkish becomes relevant: Both of these languages exhibit vowel harmony, where the vowels in many suffixes vary depending on the vowels of the root, even if consonants intervene. Thus I speculate that at least some of the reduced performance in Turkish and Finnish is due to the system not being able to recognize variants of the same suffixes as the same, and, in addition, not being able to isolate all of the roots.

Of course, in some cases, one language may represent, in some objective sense, a harder problem than another. A clear example of this is English letter-to-phoneme conversion, which, as a result of the lack of transparency in English orthography, is a harder problem that letter-to-phoneme conversion in other languages. Not surprisingly, the letter-to-phoneme systems described in e.g. (Jiampojamarn et al., 2008) and (Bartlett et al., 2008) do worse on the English test data than they do on German, Dutch, or French. On the other hand, just because one language may present a harder problem than the other doesn't mean that system developers can assume that any performance differences can be explained in such a way. If one aims to create a language-independent system, then one must explore the possibility that the system includes assumptions about linguistic structure which do not hold up across all languages.

The conclusions I would like to draw from these examples are as follows: A truly language-independent system works equally well across languages. When a system that is meant to be language independent does not in fact work equally well across languages, it is likely because something about the system design is making implicit assumptions about language structure. These assumptions are typically the result of "overfitting" to the original development language(s).[3] In Sec-

[2] In Finnish, depending on the verb class, the bare root may appear in negated present tense sentences, in second-person singular imperatives, and third-person singular present tense, or not at all (Karlsson and Chesterman, 1999). In Turkish, the bare root can function as a familiar imperative, but other forms are inflected (Lewis, 1967; Underhill, 1976).

[3] Here I use the term "overfitting" metaphorically, to call out the way in which, as the developers of NLP methodology, we rely on our intuitions about the structure of the language(s) we're working with and the feedback we get by test-

tion 4, I will argue that the best way to achieve language independence is by including, rather than eschewing, linguistic knowledge.

3 Language Independence and Language Representation at ACL

This section reports on a survey of the 119 long papers from ACL2008:HLT. Of these 119 papers, 18 explicitly claimed (16) or suggested (2) that the methods described could be applied to other languages. Another 13 could be read as implicitly claiming that. Still others present the kind of methodology that often is claimed to be cross-linguistically applicable, such as statistical machine translation. Of the 16 explicitly claiming language independence, 7 evaluated their systems on multiple languages. Since many of the techniques are meant to be cross-linguistically applicable, I collected information about the languages studied in all 119 papers. Table 1 groups the papers by how many languages (or language pairs) they study. The three papers studying zero languages involved abstract, formal proofs regarding, e.g., grammar formalisms. 95 of the papers studied just one language or language pair.

Languages or language pairs considered	Number of papers
0	3
1	95
2	13
3	3
4	2
5	1
12	1
13	1
Total	119

Table 1: Number of languages/language pairs considered

The two papers looking at the widest variety of languages were (Ganchev et al., 2008) and (Nivre and McDonald, 2008). Ganchev et al. (2008) explore whether better alignments lead to better translations, across 6 language pairs, in each direction (12 MT systems), collecting data from a variety of sources. Nivre and McDonald (2008) present an approach to dependency parsing which integrates graph-based and transition-based methods, and evaluate the result against the 13 datasets provided in the CoNLL-X shared task (Nivre et al., 2007).

It is encouraging to see such use of multilingual datasets; the field as a whole will be in a better position to test (and improve) the cross-linguistic applicability of various methods to the extent that more such datasets are produced. It is worth noting, however, that the sheer number of languages tested is not the only important factor: Because related languages tend to share typological properties, it is also important to sample across the known language *families*.

Tables 2 and 3 list the languages and language pairs studied in the papers in the survey. Table 2 presents the data on methodologies that involve producing results for one language at a time, and groups the languages by genus and family (according to the classification used by the World Atlas of Language Structures Online[4]). Table 3 presents the data on methodologies that involve symmetrical (e.g., bilingual lexicon extraction) or asymmetrical (e.g., MT) language pairs.[5]

The first thing to note in these tables is the concentration of work on English: 63% of the single-language studies involved English, and all of the language pairs studied included English as one member. In many cases, the authors did not explicitly state which language they were working on. That it was in fact English could be inferred from the data sources cited, in some cases, or from the examples used, in others. The common practice of not explicitly stating the language when it is English would seem to follow from a general sense that the methods should be crosslinguistically applicable.

The next thing to note about these tables is that many of the languages included are close relatives of each other. Ethnologue[6] lists 94 language families; ACL2008:HLT papers studied six. Of course, the distribution of languages (and perhaps more to the point, speakers) is not uniform across lan-

ing our ideas against particular languages.

[4] http://wals.info (Haspelmath et al., 2008); Note that Japanese is treated as a language isolate and Chinese is the name for the genus including (among others) Mandarin and Cantonese.

[5] The very interesting study by Snyder and Barzilay (2008) on multilingual approaches to morphological segmentation was difficult to classify. Their methodology involved jointly analyzing two languages at a time in order to produce morphological segmenters for each. Since the resulting systems were monolingual, the data from these studies are included in Table 2.

[6] http://www.ethnologue.com/ethno_docs/distribution.asp, accessed on 6 February 2009.

Language	Studies		Genus	Studies		Family	Studies	
	N	%		N	%		N	%
English	81	63.28	Germanic	91	71.09	Indo-European	109	85.16
German	5	3.91						
Dutch	3	2.34						
Danish	1	0.78						
Swedish	1	0.78						
Czech	3	2.34	Slavic	8	6.25			
Russian	2	1.56						
Bulgarian	1	0.78						
Slovene	1	0.78						
Ukranian	1	0.78						
Portuguese	3	2.34	Romance	8	6.25			
Spanish	3	2.34						
French	2	1.56						
Hindi	2	1.56	Indic	2	1.56			
Arabic	4	3.13	Semitic	9	7.03	Afro-Asiatic	9	7.03
Hebrew	4	3.13						
Aramaic	1	0.78						
Chinese	5	3.91	Chinese	5	3.91	Sino-Tibetan	5	3.91
Japanese	3	2.34	Japanese	3	3.24	Japanese	3	3.24
Turkish	1	0.78	Turkic	1	0.78	Altaic	1	0.78
Wambaya	1	0.78	West Barkly	1	0.78	Australian	1	0.78
Total	128	100.00		128	100.00		128	100.00

Table 2: Languages studied in ACL 2008 papers, by language genus and family

Source	Target	N	Source	Target	N	Symmetrical pair	N
Chinese	English	9	English	Chinese	2	English, Chinese	3
Arabic	English	5	English	Arabic	2	English, Arabic	1
French	English	2	English	French	2	English, French	1
Czech	English	1	English	Czech	2	English, Spanish	1
Finnish	English	1	English	Finnish	1		
German	English	1	English	German	1		
Italian	English	1	English	Italian	1		
Spanish	English	1	English	Spanish	1		
			English	Greek	1		
			English	Russian	1		

Table 3: Language pairs studied in ACL 2008 papers

Language family	Living lgs.	Examples	% pop.
Indo-European	430	Welsh Pashto Bengali	44.78
Sino-Tibetan	399	Mandarin Sherpa Burmese	22.28
Niger-Congo	1,495	Swahili Wolof Bissa	6.26
Afro-Asiatic	353	Arabic Coptic Somali	5.93
Austronesian	1,246	Bali Tagalog Malay	5.45
Total	3,923		84.7

Table 4: Six most populous language families, from Ethnologue

guage families. Table 4 gives the five most populous language families, again from Ethnologue.[7] These language families together account for almost 85% of the world's population.

Of course, language independence is not the only motivation for machine-learning approaches to NLP. Others include scaling to different genres within a language, robustness in the face of noisy input, the argument (in some cases) that creating or obtaining training data is cheaper than creating a rule-based system, and the difficulty in certain tasks of creating rule-based systems. Nonetheless, to the extent that language independence is an important goal, the field needs to improve both its testing of language independence and its sampling of languages to test against.

4 Linguistic Knowledge

Typically, when we think of linguistic knowledge-based NLP systems, what comes to mind are complicated, intricate sets of language-specific rules. While I would be the last to deny that such systems can be both linguistically interesting and the best approach to certain tasks, my purpose here is to point out that there are other kinds of linguistic knowledge that can be fruitfully incorporated into NLP systems. In particular, the results of language typology represent a rich source of knowledge that, by virtue of being already produced by the typologists, can be relatively inexpensively incorporated into NLP systems.

Linguistic typology is an approach to the scientific study of language which was pioneered in its modern form by Joseph Greenberg in the 1950s and 1960s (see e.g. Greenberg, 1963).[8] In the intervening decades, it has evolved from a search for language universals and the limits of language variation to what Bickel (2007) characterizes as the study of "what's where why". That is, typologists are interested in how variations on particular linguistic phenomena are distributed throughout the world's languages, both in terms of language families and geography, and how those distributions came to be the way they are.

For the purposes of improving language-independent NLP systems, we are primarily concerned with "what" and "where": Knowing "what" (how languages can vary) allows us to both broaden and parameterize our systems. Knowing "where" also helps with parameterizing, as well as with selecting appropriate samples of languages to test the systems against. We can broaden them by studying what typologists have to say about our initial development languages, and identifying those characteristics we might be implicitly relying on. This is effectively what Bilmes and Kirchhoff (2003) did in generalizing n-gram language models to factored language models. We can parameterize our systems by identifying and specifically accommodating relevant language types ("what") and then using databases produced by typologists to map specific input languages to types ("where").

The practical point of language independence is not to be able to handle in principle any possible language in the universe (human or extraterrestrial!), but to improve the scalability of NLP technology across the existing set of human languages. There are approximately 7,000 languages spoken today, of which 347 have more than 1 million speakers.[9] An NLP system that uses different parameters or algorithms for each one of a set

[7] Ibid. Example languages are included to give the reader a sense of where these language families are spoken, and are deliberately chosen to represent the breadth of each language family while still being relatively recognizable to the EACL audience.

[8] See (Ramat, to appear) for discussion of much earlier approaches.

[9] http://wwww.ethnologue.com/ethno_docs/distribution.asp; accessed 6 February 2009

of known languages is not language independent. One that uses different parameters or even algorithms for different language *types*, and includes as a first step the classification of the input language, either automatically or with reference to some external typological database, *is* language independent, at least on the relevant, practical sense.

The preeminent typological database among those which are currently publicly available is WALS: The World Atlas of Linguistic Structures Online (Haspelmath et al., 2008). WALS currently includes studies of 142 chapters studying linguistic features, each of which defines a dimension of classification, describes values along that dimension, and then classifies a large sample of languages. It is also possible to view the data on a language-by-language basis. These chapters represent concise summaries, as well as providing pointers into the relevant literature for more information.

To give a sense of how this information might be of relevance to NLP or speech systems, here is a brief overview of three chapters:

Maddieson (2008) studies tone, or the use of pitch to differentiate words or inflectional categories. He classifies languages into those with no tone systems, those with simple tone systems (a binary contrast between high and low tone), and those with more complex tone systems (more than two tone types). Nearly half of the languages in the sample have some tone, and Maddieson points out that the sample in fact underestimates the number of languages with tone.

Dryer (2008b) investigates prefixing and suffixing in inflectional morphology, looking at 10 common types of affixes (from case affixes on nouns to adverbial subordinator affixes on verbs), and using them to classify languages in terms of tendencies towards prefixing or suffixing.[10] His resulting categories are: little affixation, strongly suffixing, weakly suffixing, equal prefixing and suffixing, weakly prefixing, and strongly prefixing. The most common category (382/894 languages) is predominantly suffixing.

Dryer (2008a) investigates the expression of clausal negation. One finding of note is that all languages studied use dedicated morphemes to express negation. This contrasts with the expression of yes-no questions which can be handled with word order changes, intonation, or no overt mark at all. The types of expression of clausal negation that Dryer identifies are: negative affix, negative auxiliary verb, and negative particle. In addition, some languages are classified as using a negative word that may be a verb or may be a particle, as having variation between negative affixes and negative words, and as having double (or two-part) negation, where each negative clause requires two markers, one before the verb, and one after it.

These examples illustrate several useful aspects of the knowledge systematized by linguistic typology: First, languages show variation beyond that which one might imagine looking only at a few familiar (and possibly closely related) languages. Second, however, that variation is still bounded: Though typologists are always interested in finding new categories that stretch the current classification, for the purposes of computational linguistics, we can get very far by assuming the known types exhaust the possibilities. Finally, because of the work done by field linguists and typologists, this knowledge is available as high-level generalizations about languages, of the sort that can inform the design of linguistically-sophisticated, language-independent NLP systems.

5 Conclusion

This paper has briefly argued that the best way to create language-independent systems is to include linguistic knowledge, specifically knowledge about the ways in which languages vary in their structure. Only by doing so can we ensure that our systems are not overfitted to the development languages. Furthermore, this knowledge is relatively inexpensive to incorporate, as it does not require building or maintaining intricate rule systems. Finally, if the field as a whole values language independence as a property of NLP systems, then we should ensure that the languages we select to use in evaluations are representative of both the language types and language families we are interested in.

Acknowledgments

I am grateful to Stephan Oepen and Timothy Baldwin for helpful discussion. Any remaining infelicities are my own. This material is based in part upon work supported by the National Science Foundation under Grant No. 0644097. Any opinions, findings, and conclusions or recommenda-

[10] For the purposes of this study, he sets aside less common inflectional strategies such as infixing, tone changes, and stem changes.

tions expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2008. Automatic syllabification with structured SVMs for letter-to-phoneme conversion. In *Proceedings of ACL-08: HLT*, pages 568–576, Columbus, Ohio, June. Association for Computational Linguistics.

Balthasar Bickel. 2007. Typology in the 21st centure: Major current developments. *Linguistic Typology*, pages 239–251.

Jeff A. Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *in Proceedings of HLT/NACCL, 2003*, pages 4–6.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 155–163, Rochester, New York, April. Association for Computational Linguistics.

Matthew S. Dryer. 2008a. Negative morphemes. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. Available online at http://wals.info/feature/112. Accessed on 2009-02-07.

Matthew S. Dryer. 2008b. Prefixing vs. suffixing in inflectional morphology. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. Available online at http://wals.info/feature/26. Accessed on 2009-02-07.

Kuzman Ganchev, João V. Graça, and Ben Taskar. 2008. Better alignments = better translations? In *Proceedings of ACL-08: HLT*, pages 986–993, Columbus, Ohio, June. Association for Computational Linguistics.

Joseph Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In *Univerals of Language*, pages 73–113. MIT Press, Cambridge.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors. 2008. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. http://wals.info.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint processing and discriminative training for letter-to-phoneme conversion. In *Proceedings of ACL-08: HLT*, pages 905–913, Columbus, Ohio, June. Association for Computational Linguistics.

Fred Karlsson and Andrew Chesterman. 1999. *Finnish: An Essential Grammar*. Routledge, London.

Geoffrey Lewis. 1967. *Turkish Grammar*. Clarendon Press, Oxford.

Ian Maddieson. 2008. Tone. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. Available online at http://wals.info/feature/13. Accessed on 2009-02-07.

Joakim Nivre and Ryan McDonald. 2008. Integrating graph-based and transition-based dependency parsers. In *Proceedings of ACL-08: HLT*, pages 950–958, Columbus, Ohio, June. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic, June. Association for Computational Linguistics.

Paolo Ramat. to appear. The (early) history of linguistic typology. In *The Oxford Handbook of Linguistic Typology*. Oxford University Press, Oxford.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio, June. Association for Computational Linguistics.

Robert Underhill. 1976. *Turkish Grammar*. MIT Press, Cambridge, MA.