CS11-737: Multilingual Natural Language Processing

Typology: The Space of Languages

Yulia Tsvetkov
Linguistic diversity: ~7000 languages
Low resource languages

There are about 460 languages in India.
1.38 billion people
Low resource languages

Africa is a continent with a very high linguistic diversity: there are an estimated 1.5-2K African languages from 6 language families. 1.33 billion people
Low-resource/multilingual NLP

40% of world’s population: South Asia - 1.75 billion, Africa - 1.3 billion, etc.
Approaches to low-resource/multilingual NLP

- Manual curation and annotation of large-scale resources for thousands of languages in infeasible or prohibitively expensive.

- Unsupervised learning (Snyder and Barzilay 2008; Cohen and Smith, 2009; Snyder, 2010; Vulić, De Smet, and Moens 2011; Spitkovsky et al., 2011; Goldwasser et al., 2011; Titov and Klementiev 2012; Baker et al., 2014, and many others)
Approaches to low-resource/multilingual NLP

● Cross-lingual transfer learning – transfer of resources and models from resource-rich source to resource-poor target languages
  ○ Transfer of annotations (e.g., POS tags, syntactic or semantic features) via cross-lingual bridges (e.g., word or phrase alignments)
  ○ Transfer of models – train a model in a resource-rich language and adapt (e.g. fine-tune) it in a resource-poor language

● Zero-shot learning – train a model in one domains and assume it generalizes more or less out-of-the-box in a low-resource domain

● Few shot learning – train a model in one domain and use only few examples from a low-resource domain to adapt it
● Joint multilingual learning – train a single model on a mix of datasets in all languages, to enable data and parameter sharing where possible
Choosing transfer languages

How to define similarity across languages?

- Word overlap and sub-word overlap
  - Russian – Русский
  - Ukrainian – Українська
  - Chinese – 中文
  - Korean – 한국어
  - Vietnamese – Tiếng Việt
  - Georgian – ქართული
  - Japanese – 日本人
  - Turkish – Türk
  - Hebrew – עברית
  - Arabic – عربي
  - Hindi – हिन्दी
  - Xhosa – isiXhosa

- Areal similarity [www.glottolog.org]

- Demographic similarity
Genealogical similarity

1. Niger–Congo (1,542 languages) (21.7%)
2. Austronesian (1,257 languages) (17.7%)
3. Trans–New Guinea (482 languages) (6.8%)
4. Sino-Tibetan (455 languages) (6.4%)
5. Indo-European (448 languages) (6.3%)
6. Australian [dubious] (381 languages) (5.4%)
7. Afro-Asiatic (377 languages) (5.3%)
8. Nilo-Saharan [dubious] (206 languages) (2.9%)
9. Oto-Manguean (178 languages) (2.5%)
10. Austroasiatic (167 languages) (2.3%)
11. Tai–Kadai (91 languages) (1.3%)
12. Dravidian (86 languages) (1.2%)
13. Tupian (76 languages) (1.1%)
Typological similarity

- Linguistic typology: classification of languages according to their functional and structural properties
  - explains common properties across languages
  - explains structural diversity across languages

“The classification of languages or components of languages based on shared formal characteristics.”
Linguistic typology example: phonology

<table>
<thead>
<tr>
<th>Place</th>
<th>Bilabial</th>
<th>Labio-lingual</th>
<th>Linguo-labial</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Palato-alveolar</th>
<th>Retroflex</th>
<th>Alveolo-palatal</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td>m</td>
<td>m</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Stop</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>t</td>
<td>q</td>
<td>c</td>
<td>k</td>
<td>q</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Sibilant affricate</td>
<td>ts dz</td>
<td>tf dz</td>
<td>ts dz</td>
<td>ts dz</td>
<td>ts dz</td>
<td>ts dz</td>
<td>ts dz</td>
<td>ts dz</td>
<td>ts dz</td>
<td>ts dz</td>
<td>ts dz</td>
<td>ts dz</td>
<td>ts dz</td>
</tr>
<tr>
<td>Non-sibilant affricate</td>
<td>pθ β pθ β</td>
<td>pθ β pθ β</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
</tr>
<tr>
<td>Sibilant fricative</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
<td>s z j z</td>
</tr>
<tr>
<td>Non-sibilant fricative</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
<td>θ θ dθ</td>
</tr>
<tr>
<td>Approximant</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
<td>ɣ γ j γ</td>
</tr>
<tr>
<td>Flap or tap</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
<td>j j j</td>
</tr>
<tr>
<td>Trill</td>
<td>B</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>Lateral affricate</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
<td>tθ dθ</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
<td>tθ j θ</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
</tr>
<tr>
<td>Lateral flap</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
<td>ɣ γ</td>
</tr>
</tbody>
</table>
Linguistic typology example: numerals

Feature 131A: Numeral Bases
2,676 languages, 192 attributes

<table>
<thead>
<tr>
<th>ID#</th>
<th>Feature Name</th>
<th>Category</th>
<th>Feature Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Consonant Inventories</td>
<td>Phonology (19)</td>
<td>{1: Large, 2: Small, 3: Moderately Small, 4: Moderately Large, 5: Average}</td>
</tr>
<tr>
<td>23</td>
<td>Locus of Marking in the Clause</td>
<td>Morphology (10)</td>
<td>{1: Head, 2: None, 3: Dependent, 4: Double, 5: Other}</td>
</tr>
<tr>
<td>30</td>
<td>Number of Genders</td>
<td>Nominal Categories (28)</td>
<td>{1: Three, 2: None, 3: Two, 4: Four, 5: Five or More}</td>
</tr>
<tr>
<td>58</td>
<td>Obligatory Possessive Inflection</td>
<td>Nominal Syntax (7)</td>
<td>{1: Absent, 2: Exists}</td>
</tr>
<tr>
<td>66</td>
<td>The Perfect</td>
<td>Verbal Categories (16)</td>
<td>{1: None, 2: Other, 3: From ‘finish’ or ‘already’, 4: From Possessive}</td>
</tr>
<tr>
<td>81</td>
<td>Order of Subject, Object and Verb</td>
<td>Word Order (17)</td>
<td>{1: SVO, 2: SOV, 3: No Dominant Order, 4: VSO, 5: VOS, 6: OV, 7: OSV}</td>
</tr>
<tr>
<td>121</td>
<td>Comparative Constructions</td>
<td>Simple Clauses (24)</td>
<td>{1: Conjoined, 2: Locational, 3: Particle, 4: Exceed}</td>
</tr>
<tr>
<td>125</td>
<td>Purpose Clauses</td>
<td>Complex Sentences (7)</td>
<td>{1: Balanced/deranked, 2: Deranked, 3: Balanced}</td>
</tr>
<tr>
<td>138</td>
<td>Tea</td>
<td>Lexicon (10)</td>
<td>{1: Other, 2: Derived from Sinitic ‘cha’, 3: Derived from Chinese ‘te’}</td>
</tr>
<tr>
<td>140</td>
<td>Question Particles in Sign Languages</td>
<td>Sign Languages (2)</td>
<td>{1: None, 2: One, 3: More than one}</td>
</tr>
<tr>
<td>142</td>
<td>Para-Linguistic Usages of Clicks</td>
<td>Other (2)</td>
<td>{1: Logical meanings, 2: Affective meanings, 3: Other or none}</td>
</tr>
</tbody>
</table>

Example from Georgi, Xia and Lewis (2010)

Automatic prediction of typological features

- Morphosyntactic annotation projection
  - Sentence and treebank alignments to project feature annotations from similar languages

- Unsupervised and semi-supervised feature propagation
  - Hierarchical typological clustering and majority value assignment
  - Language-family based nearest neighbor projection
  - Matrix completion

- Supervised Learning
  - Logistic regression
  - Determinant point process with neural features

- Cross-lingual distributional feature alignment


TyP-NLP Workshop at ACL 2019
<table>
<thead>
<tr>
<th>Name</th>
<th>Levels</th>
<th>Coverage</th>
<th>Feature Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>World Atlas of Language Structures (WALS)</td>
<td>Phonology, Morphosyntax, Lexical semantics</td>
<td>2,676 languages; 192 attributes; 17% values covered</td>
<td>ORDER OF OBJECT AND VERB Amele: OV (713) Gbaya Kara: VO (725)</td>
</tr>
<tr>
<td>Atlas of Pidgin and Creole Language Structures (APICS)</td>
<td>Phonology, Morphosyntax</td>
<td>76 languages; 355 attributes</td>
<td>TENSE–ASPECT SYSTEMS: Ternate Chabacano: purely aspectual (10) Afrikaans: purely temporal (1)</td>
</tr>
<tr>
<td>UERL, Typological Compendium</td>
<td>Phonology, Morphosyntax,Lexical semantics</td>
<td>8,070 languages; 284 attributes; 439,000 values</td>
<td>CASE IS PREFIX Berber (Middle Atlas): yes (38) Hawaiian: no (993)</td>
</tr>
<tr>
<td>Syntactic Structures of the World’s Languages (SSWL)</td>
<td>Morphosyntax</td>
<td>262 languages; 148 attributes; 45% values covered</td>
<td>STANDARD NEGATION IS SUFFIX Amharic: yes (21) Laal: no (170)</td>
</tr>
<tr>
<td>AUTOTYP</td>
<td>Morphosyntax</td>
<td>825 languages; 1,000 attributes</td>
<td>PRESENCE OF CLUSIVITY ‘Kung (Ju): false Ik (Kulua): true</td>
</tr>
<tr>
<td>Valency Patterns Leipzig (ValPaL)</td>
<td>Predicate–argument structures</td>
<td>36 languages; 80 attributes; 1,156 values</td>
<td>TO LAUGH Mandinka: 1 &gt; V Shannon: Vab[1] 1</td>
</tr>
<tr>
<td>Lyon–Albuquerque Phonological Systems Database (LAIPsyD)</td>
<td>Phonology</td>
<td>422 languages; 70 attributes</td>
<td>cf. AND T Sindihi: yes (1) Chuvash: no (421)</td>
</tr>
<tr>
<td>PHOIBLE Online</td>
<td>Phonology</td>
<td>2,155 languages; 2,160 attributes</td>
<td>m Vietnamese: yes (2053) Firiha: no (152)</td>
</tr>
<tr>
<td>StressTyp2</td>
<td>Phonology</td>
<td>699 languages; 927 attributes</td>
<td>STRESS ON FIRST SYLLABLE Koromfè: yes (183) Cubeo: no (516)</td>
</tr>
<tr>
<td>World Loanword Database (WOLD)</td>
<td>Lexical semantics</td>
<td>41 languages; 24 attributes; 2,000 values</td>
<td>HORSE Quechua: Kabulla bornewed (24) Sakha: ałg̲a no evidence (18)</td>
</tr>
<tr>
<td>Intercontinental Dictionary Series (IDS)</td>
<td>Lexical semantics</td>
<td>329 languages; 1,310 attributes</td>
<td>WORLD Russian: mir Tocharian A: arkişi</td>
</tr>
<tr>
<td>Automated Similarity Judgment Program (ASJP)</td>
<td>Lexical semantics</td>
<td>7,221 languages; 40 attributes</td>
<td>I Ainu Masuko: co'oluy Japanese: naunashi</td>
</tr>
</tbody>
</table>

URIEL

- URIEL typological compendium
  - Phonology, morphosyntax, lexical semantics
  - 8,070 languages, 284 attributes, $439,000 values
- lang2vec representations from URIEL
  https://pypi.org/project/lang2vec/


Linguistic universals

- All languages have vowels and consonants
- All (or at least nearly all) languages of the world also make a distinction between nouns and verbs
Linguistic typology in NLP

Open research problems

● how to extract typological features automatically from existing multilingual resources such as Universal Dependency treebank, UniMorph, Wikipedia, or Bible corpora
● how to accurately predict typological knowledge while controlling for genealogical and areal biases
● how to incorporate linguistic typology into models
● how to alleviate negative transfer and catastrophic forgetting in multilingually trained models using typological knowledge
Further readings

● Survey:


● Papers in tracks on morphology/phonology or multilinguality at *CL conferences

● Workshops: SIGMORPHON, SIGTYP, ComputEL, AfricaNLP, DeepLo, etc.
Class reading and discussion

● Reading

● Discussion question