Silviu Cucerzan & David Yarowsky & Richard Wicentowski (2002 and 2003)

NPFL128 - Language Technologies in Practice

Adriana Rodríguez Flórez (adrirflorez@gmail.com)

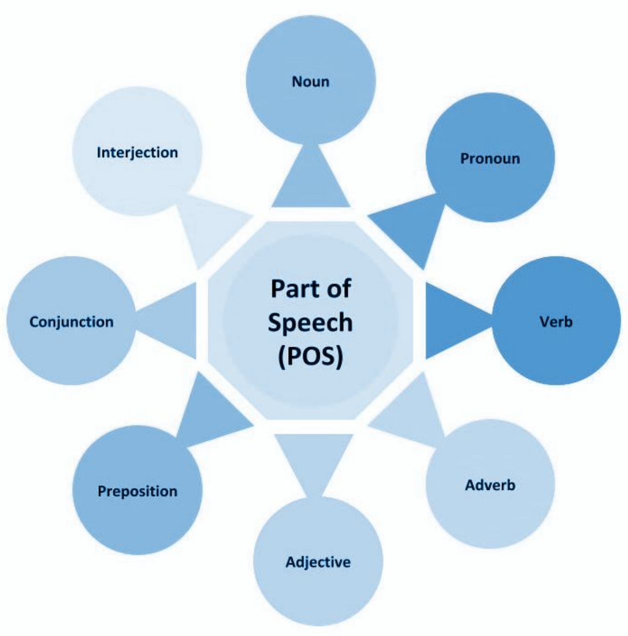
Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day

Silviu Cucerzan and David Yarowsky
Department of Computer Science and
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218 USA
{silviu,yarowsky}@cs.jhu.edu

Abstract

This paper presents a method for bootstrapping a fine-grained, broad-coverage part-of-speech (POS) tagger in a new language using only one person-day of data acquisition effort. It requires only three resources, which are currently readily available in 60-100 world languages: (1) an online or hard-copy pocket-sized bilingual dictionary, (2) a basic library reference grammar, and (3) access to an existing monolingual text corpus in the language. The algorithm begins by inducing initial lexical POS distributions from English translations in a bilingual dictionary without POS tags. It handles irregular, regular and semi-regular morphology through a robust generative model using weighted Levenshtein alignments. Unsupervised induction of grammatical gender is performed via global modeling of context-window feature agreement. Using a combination of these and other evidence sources, interactive training of context and lexical prior models are accomplished for fine-grained POS tag spaces. Experiments show high accuracy, fine-grained tag resolution with minimal new human effort.

1. Multilingual PoS Tagger



- PoS tagging with minimal effort

- But... what's *minimal*?
  - partially-tagged corpora?
  - small seed inputs?
  - re-using annotated data trans-lingually?
  - ittle human and resource costs?

- Minimal supervision via use of existing readily available basic resources

  1. Bilingual dictionary
  2. Reference grammar
  3. Monolingual text corpus

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day (Cucerzan & Yarowsky, 2002). Johns Hopkins University.

1

Extract basic PoS distributions from unlabeled bilingual dictionaries

2

Expand inflectional forms via morphological models

3

Use contextual learning to refine PoS tagging incl. grammar features

4

Re-train and refine iteratively for better accuracy

Total time: 1-person day of work

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day (Cucerzan & Yarowsky, 2002). Johns Hopkins University.

1.1. Inducing PoS tags from unlabeled bilingual dictionaries

Romanian	True POS	English translation list
mandat	N	warrant; proxy; mandate; money order; power of attorney
manechin	N	model, dummy
manifesta	V	arise, express itself, show
manual	Adj	manual;
	N	manual; textbook; handbook
mare	Adj	large; big; great; tall; old; important;
	N	sea
maro	Adj	brown, chestnut

Figure 1: A sample Romanian-English dictionary. The POS tags are used only for evaluation and are not available in many bilingual dictionaries.

- Process: foreign word → English translation → estimate probabilities → induce PoS tag
- Assumption: PoS of English translations is consistent cross-linguistically, for phrases

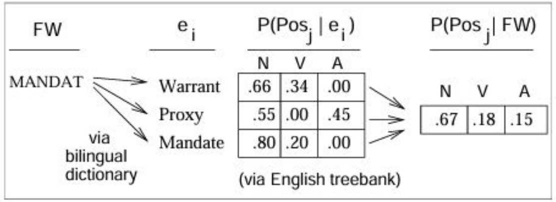


Figure 2: Inducing a preliminary POS distribution for the Romanian word *mandat* via a simple English translation list.

$$P(T_f|w_{e_1}...w_{e_n}) = P(T_f|T_{e_1}...T_{e_n}) \cdot P(T_{e_1}...T_{e_n}|w_{e_1}...w_{e_n})$$

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day (Cucerzan & Yarowsky, 2002). Johns Hopkins University.

- Key: estimate a robust tag probability distribution with large enough probability for the true PoS, so that we can seed further training with these minimal resources

Target Language	Training Dictionary	Accuracy Exact POS	Correct POS Over Threshold	Coverage	Mean Probability of Truth
Romanian	Spanish - English	92.9	97.8	98	.91
Kurdish	Spanish - English	76.8	93.1	95	.82
Spanish	Romanian - English	83.3	94.9	97	.86

Table 1: Performance of inducing candidate part-of-speech distributions derived solely from untagged English translation lists. Results are measured by type (all dictionary entries are weighted equally).

- Errors from:
  - Differing annotating/formatting styles across dictionaries
  - Untagged English translations (rare or proper nouns)
  - OCR failure in resource acquisition
  - Equal weights for all words, incl. extremely rare ones
  - Ambiguity in PoS tags and definitions

1.2. Inducing morphological analyses from ref. grammars

Root Affix	Inflected Affix	Part-of-speech Tag
Spanish:		
o$	o$	Adj-masc-sing
o$	os$	Adj-masc-plur
o$	a$	Adj-fem-sing
o$	as$	Adj-fem-plur
e$	e$	Adj-masc,fem-sing
e$	es$	Adj-masc,fem-plur
ar$	o$	Verb-Indic_Pres-p1-sing
ar$	as$	Verb-Indic_Pres-p2-sing
ar$	a$	Verb-Indic_Pres-p3-sing
ar$	amos$	Verb-Indic_Pres-p1-plur
ar$	áis$	Verb-Indic_Pres-p2-plur
ar$	an$	Verb-Indic_Pres-p3-plur

Table 2: Sample extracted regular inflectional paradigms (suffix context is marked by $).

- Process: creation of inflectional affix tables → weighted Levenshtein distances within a corpus
- Manually entering inflectional paradigms based on reference grammars for a language, ca. 200 lines
  - Incl. closed class words (very short or extremely rare ones that would not generalize well)

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j)=0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j)+1 \\ \operatorname{lev}_{a,b}(i,j-1)+1 \\ \operatorname{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

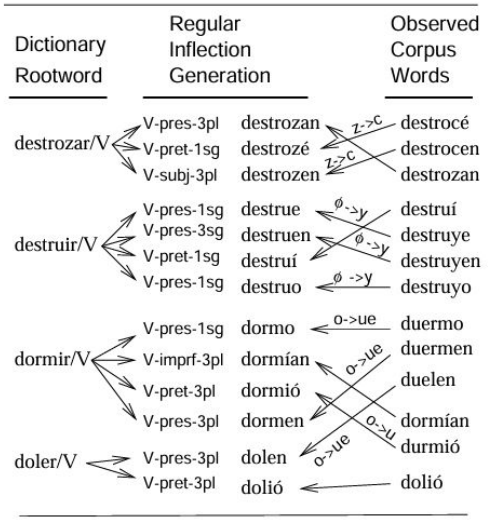


Figure 3: Inflectional analysis induction via weighted string alignment to noisy generations from dictionary roots under regular paradigms

- Takes into account potential irregularities and stem changes for more morphologically complex languages!
    - Even if stem changes are not explicitly accounted for in created tables… :-)
    - Not really for close-classed words (hence manual incl.) :-(

weighted mixture model

- System combines similar and pseudo-regular generated words PoS/morphological induced data to improve prediction of the next word
    - Based on relevance, multiple possible tags assigned
    - Artificial word forms that have been created by applying common morphological rules (like suffix changes) but may still contain irregularities

1.3. PoS model induction and PoS subtags

- Suffix-based modelling with trie smoothing
- Paradigmatic cross-context tag modelling for larger corpora
- Contextual agreement for sub-PoS grammar features

- Assumption: same PoS words usually occur in similar syntactic environments and somewhat narrow contextual windows

- \+ we must also have enough of PoS/ morphological instances for each

Agreement window assumption also holds for sub grammatical features like gender

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day (Cucerzan & Yarowsky, 2002). Johns Hopkins University.

1.4. Induction of gender with contextual learning

- Assumption: ike PoS, Certain grammatical features tend to occur in a given contextual window

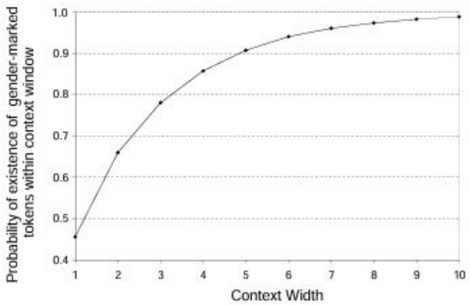


Figure 5: The probability that at least one gender-marked word will occur within a window of $\pm i$ words relative to another gender marked word (of any part of speech).

Tested over potential window sizes and 3 yielded the best overall coverage and accuracy

$$P(Gen_k|w) = \frac{1}{N} \sum_{i \in loc(w)} \sum_{j=-3}^{+3} P(Gen_k|w_{i+j}) Wt(j)$$

- Combining aforementioned models with this contextual window allows to infer gender for words and suffixes alike
  - Not so good for Spanish tho :-(

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day (Cucerzan & Yarowsky, 2002). Johns Hopkins University.

(a small break from the PoS tagger itself to delve more into…)

Minimally Supervised Induction of Grammatical Gender

Silviu Cucerzan and David Yarowsky
Department of Computer Science and
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218, USA
{silviu,yarowsky}@cs.jhu.edu

Abstract

This paper investigates the problem of determining grammatical gender for the nouns of a language starting with minimal resources: a very small list of seed nouns for which gender is known or via translingual projection of natural gender. We show that through a bootstrapping process that uses contextual clues from an unannotated corpus and morphological clues modeled with suffix tries, accurate gender predictions can be induced for five diverse test languages.

2. Induction of grammatical gender

feminine		masculine	
A:	la casa, la cara, la mesa, la cama, la silla, la cerveza	O:	el carro, el dinero, el florero, el edificio
CIÓN: SIÓN:	la canción, la relación la presión, la televisión	AJE:	el mensaje, el paisaje, el garaje, el pasaje
DAD: TAD:	la edad, la verdad la amistad, la lealtad	OR:	el amor, el dolor, el error, el sabor, el temor
IRREGULAR:	la foto, la mano, la moto, la radio	IRREGULAR:	el clima, el día, el idioma, el poema

- Grammatical gender assignment with <u>minimal effort</u>
- Again, minimization via readily via use of existing <u>readily available basic resources</u>
- <u>Gender:</u> Intrinsic property of nouns found in many languages, but...
- Culture-specific and arbitrary
    - Fem/Masc, or combined?
    - Neut, ambiguous with masc?
    - (In)Animacy
    - Natural sex or to morphological rules?

Minimally Supervised Induction of Grammatical Gender (Cucerzan & Yarowsky, 2003). Johns Hopkins University.

2.1. Seeding gendered nouns

1

Extract list of seed nouns from unlabeled bilingual dictionaries

2

Frequency-based context analysis for all gender classes found in corpus

3

Morphological disambiguation via suffix-based patterns

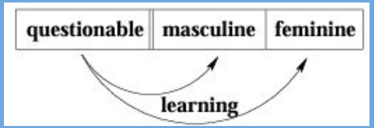
- Manually select high confidence, translingual:
    - Specifically nouns! Disambiguate homonymous forms
    - English natural gender → target natural gender
- Generalize from dictionary-extracted seed wordlist:
    - Frequency, suffix-based patterns + Contextual analysis

Neuter = masc?

Same context = same gender?

Minimally Supervised Induction of Grammatical Gender (Cucerzan & Yarowsky, 2003). Johns Hopkins University.

2.2. Learning gender via context



- Frequency analysis of seeded gender classes co-occurrence based on a threshold
    - Distributions re-estimated w.r.t. reliable context occurrence

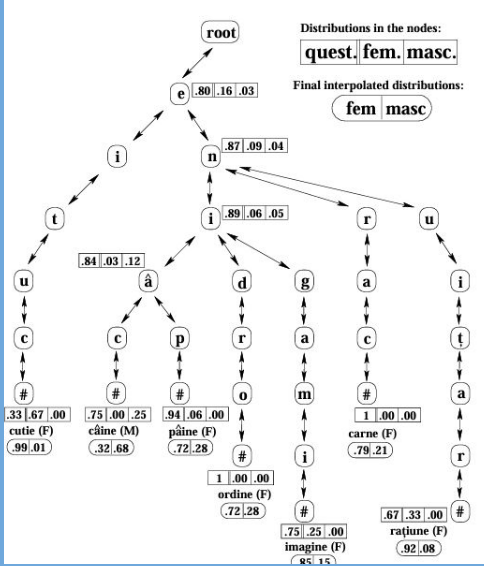
- Left, right, bilateral context for word and sub-word structures
    - Language-specific!

- Very low coverage :-(
    - Achieve high confidence for a small set of nouns
    - No valid contexts found for the greater majority

- Work-around: use morphological analysis models for disambiguation
    - Variable-length suffix patterns
    - Assignment of gender with greater suffix variability (language-spec), for unknown, un-disambiguable words

Minimally Supervised Induction of Grammatical Gender (Cucerzan & Yarowsky, 2003). Johns Hopkins University.

2.3. Suffix-based analysis in tries



- Affix modeling in trie structure → to be smoothed
  - For nouns with no reliable contexts: more aligned with nouns sharing longer suffixes

$$\widehat{P}(gen_j|l_n l_{n-1} ... l_i) = P_{node(l_n l_{n-1} ... l_i)}(gen_j) + P_{node(l_n l_{n-1} ... l_i)}(quest) \cdot \widehat{P}(gen_j|l_n l_{n-1} ... l_{i+1})$$

- Gender probability estimated from tree path recursively
  - Parametrized to weight of suffix-sharing preceding notes in the trie (α, β)
- Regarding PoS tagging...
  - True PoS shall not be ignored!
  - Presence of PoS-differing homonymous words

2.4. Gender induction results

Spanish	Natural gender seeds	(53 fem., 51 masc.)		
	by type		by token	
2993 nouns	context	+morph.	context	+morph.
coverage	54.06	100	72.71	100
accuracy	**98.70**	**95.59**	99.47	98.45
Spanish	**System extracted seeds (18 fem., 30 masc.)**			
	by type		by token	
2993 nouns	context	+morph.	context	+morph.
coverage	50.84	100	77.33	100
accuracy	**98.69**	**95.49**	99.51	98.13

Table 7: Results for Spanish

- Generalized inference of general gender rules in morphology for 5 languages

- Hindered by:
  - Absence of contextual clues
  - Language-specific exceptions and rules
  - Lack of natural gender distinction
  - Equivalence of genders in contexts

- 🌟 El agua fría :-)

Minimally Supervised Induction of Grammatical Gender (Cucerzan & Yarowsky, 2003). Johns Hopkins University.

1.5. and PoS induction results

	Spanish	Romanian	
	NNS 8h	NNS 8h	NNS-8h NS-4h
All words			
core-tag	93.1	86.3	89.2
exact-match	86.5	68.6	75.5
exact w/o gender	87.0	76.7	83.0
Nouns			
core-tag	90.3	97.4	97.4
*number	100.0	97.4	98.9
*gender	100.0	54.9	64.7
*definiteness	–	96.6	93.7
*case	–	97.4	97.4
Verbs			
core-tag	94.7	87.9	89.5
*tense	93.0	92.6	93.2
*number	100.0	91.5	91.2
*person	97.2	92.6	93.2
Adjectives			
core-tag	79.7	78.6	81.5
*gender	100.0	81.3	82.2
*number	100.0	98.3	98.3

Table 3: Performance of POS tagger induction based on 1 person-day of supervision, no tagged training corpora and a fine-grained (≈250 tags) tagset. NNS and NN refer to non-native-speaker and native-speaker effort.

- Multilingual PoS tagging (ca. 250 tagged) for Romanian and Spanish
    - Higher resource presence + knowledge of researches
- Hindered by:
    - Annotation/Formatting style differences in selected resources
    - Differing and non-existing PoS, language-spec
    - Language-specific exceptions and rules
    - Lack of natural gender distinction
    - Equivalence of genders in contexts
- Determiner or adjective? Policy or error?

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day (Cucerzan & Yarowsky, 2002). Johns Hopkins University.

Is it worth to have 4h hours of NS (Romanian) apart from 8h of NNS?

🌟Discussion fuel 🌟

- Standardization of PoS tagging policies across resources?
- Levenshtein distance use for other morphologies, i.e. template-based?
- Contextual-based analysis for flexible word-order syntax? Scalable to more understandings of gender and animacy?
- Minimal supervision:
  - To what extent?
  - For lower resource languages?
- Gender/PoS induction influenced by gender bias in resources?

Cucerzan, Silviu, & Yarowsky, David. (2002).
Bootstrapping a Multilingual Part-of-Speech Tagger in One Person-day.

Yarowsky, David. & Wicentowski, Richard. (2003).
Minimally Supervised Induction of Grammatical Gender.
Proceedings of HLT-NAACL 2003, Edmonton, Canada, pp. 40-47.

02/03/2025 - ÚFAL, Charles University (Prague)
Adriana Rodríguez Flórez (adrirflorez@gmail.com)