Frugal Paradigm Completion

Erdmann, A., Kenter, T., Becker, M., and Schallhart, C. (2020)

What's a paradigm?

- Linguistic object abstractly represented by a *lexeme* and divisible into subparts called cells
- Morphosyntactic features define cells
- Cells are filled with phonological material (forms)
- A useful formalisation is then:
  - (lemma, cell, form)
  - e.g., (SING, 3.sg.pres, *sing*)
- ...in this paper, paradigms are *not* linguistic objects, and should be thought of as 3d arrays or lists of triples

SING	present	past
3.sg	sings	sang
3.pl	sing	sang

lemma	cell	form
SING	3.sg.pres	sing+s
SING	3.sg.pst	sang
SING	3.pl.pres	sing
SING	3.pl.pst	sang

What's Paradigm Completion?

- How many lexical entries, i.e., rows/sources, do you need to provide in order to predict all remaining cells?
- This is especially problematic across inflection classes
- You can either
  - Split up the inflection classes and require multiple sources, *a priori*
  - Only sources for as long as there are errors in what they predict (even if an inflection class isn't fully disambiguated)

lexeme	cell	form
WALK	3.sg.pst	walked
FLY	3.sg.pst	
SING	3.sg.pst	

lexeme	cell	form	class
WALK	3.sg.pst	walked	+ed
FLY	3.sg.pst	flew	y > ew
SING	3.sg.pst	sang	i > a

Paradigm Completion: Inflection class disambiguation

Lemma-based Paradigm Completion (Dreyer & Eisner 2011)

- Based on a traditional conception of a paradigm (as a linguistic object that is equivalent to its lexeme)
- You don’t need to know how to find a source, because the source is always going to be the lemma
- All you need to know is the probability that a form belongs to a given lemma

Paradigm Completion: Inflection class disambiguation

Most Informative Source (Kann & Schütze 2018)

- Sources are selected based on how informative they are
- “they do not attempt to minimize the number of unique sources that must be manually realized” (Erdmann et al. 2020:8249)

Paradigm Completion: Inflection class disambiguation

Principal Parts: highly informative sources from which you can deduct other forms

Static Principal Parts (Finkel & Stump 2007; Stump & Finkel 2013)

- Use the same principal parts for every paradigm

Dynamic Principal Parts (Finkel & Stump 2007)

- Use highly representative/informative principal parts for each inflection class

Adaptive Principal Parts (Finkel & Stump 2007)

- Start with a static principal part in one paradigm, but use dynamic parts as needed if an inflection class calls for it

Frugal Paradigm Completion

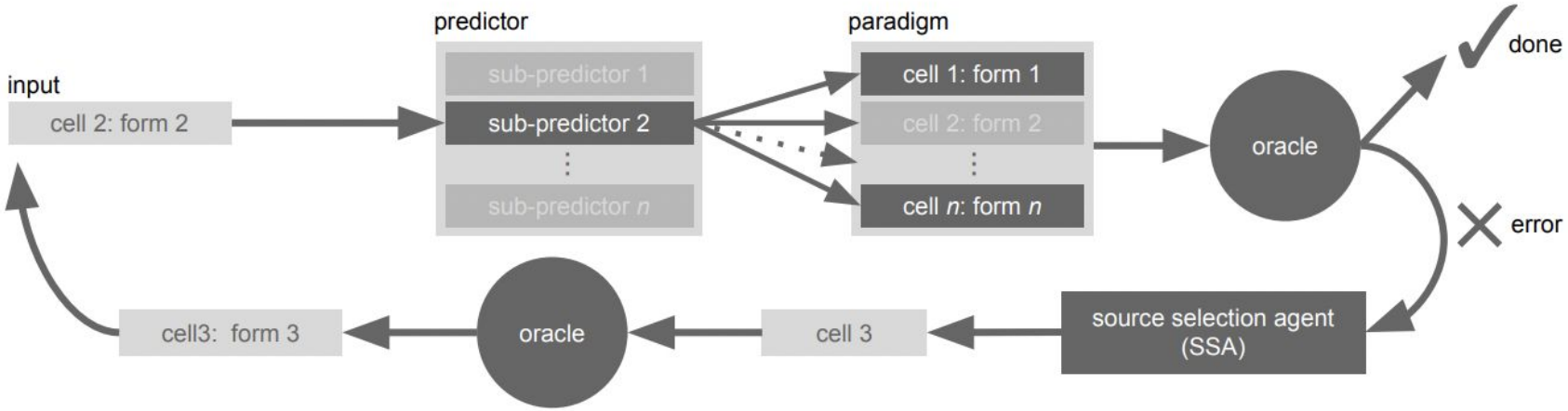

Figure 1: Schematic representation of the flow of Frugal Paradigm Completion at inference time.

Frugal Paradigm Completion

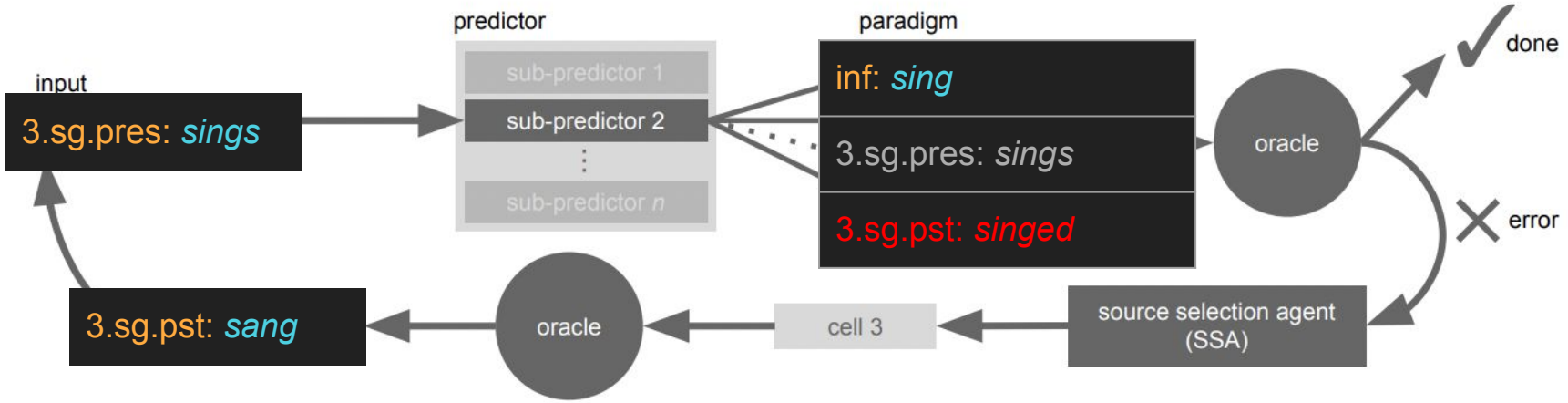

Figure 1: Schematic representation of the flow of Frugal Paradigm Completion at inference time.

Parts of the Frugal Paradigm Completion (FPC)

Predictor
- Input: a source cell **in_NFIN f l y**
- Composed of subpredictors that are trained on full paradigms to complete every cell
- Output: a target form

Source Selection Agent (SSA)
- Remembers previous cells chosen for that paradigm and their forms
- E.g., it won't request **in_PST.PTCP** if the corresponding form for **in_PST** ended in **ed**
- Trained on an idealised source selection for "minimum set covers"
- The highest ranking cell in each cover gets selected as the default source cell

Oracle
- Decides if a paradigm is correct
- Provides a source's cell-form realisation to the predictor

Experiments

The authors compared FPC against 3 other paradigm completion methods on 7 languages (Arabic, German, English, Russian, Latin, Hungarian, Irish).

Each language divided into training, testing, and development sets

Evaluation was done against their own metric (*auto-rate*), which is a proportion of the number of realisations correctly predicted such that they were NOT provided as sources (this is why accuracy alone would always be 100%).

The other methods: FPC with random SSA, the lemma-sourced approach, and static principal parts (both single- and multi-sourced).

Results

FPC wins by the autorate metric because

- it does not need to wait to establish a fully inflection class and
- the predictiveness of each source is always checked along the way.

Interestingly, FPC does not seem to be much better than FPC with randomised sources. The authors conclude that it does not seem to matter what the source is!

	Accuracy		Auto-rate		Mcpp	
	Dev	Test	Dev	Test	Dev	Test
Arabic nouns						
Lemma	62.0	58.8	59.3	56.5	9.6	10.7
Static	95.9	99.4	89.5	93.1	2.9	2.1
Random Ag.			**90.2**	90.9	**2.2**	2.2
FPC			**90.2**	**93.6**	**2.2**	**1.5**
German verbs						
Lemma	87.6	89.0	84.1	85.8	4.3	4.0
Static	94.1	96.4	86.7	88.9	3.6	3.0
Random Ag.			90.0	92.1	2.4	1.9
FPC			**91.8**	**92.5**	**2.0**	**1.8**
English verbs						
Lemma	96.5	94.0	76.7	74.2	1.2	**1.3**
Static	99.7	98.4	39.7	38.4	3.0	3.0
Random Ag.			76.0	73.3	1.2	1.4
FPC			**77.3**	**74.3**	**1.1**	**1.3**
Russian nouns						
Lemma	97.1	95.6	88.3	87.5	**1.3**	1.5
Static	98.4	98.3	72.6	72.3	3.2	3.2
Random Ag.			86.1	84.3	1.6	1.8
FPC			**88.5**	**89.1**	**1.3**	**1.2**

Latin nouns						
Lemma	65.5	51.6	63.6	49.6	5.1	6.7
Static	97.7	96.8	80.8	79.7	2.3	2.4
Random Ag.			85.9	84.7	1.7	1.8
FPC			**89.0**	**87.8**	**1.3**	**1.4**
Hungarian nouns						
Lemma	95.6	90.9	92.8	88.0	2.5	4.1
Random Ag.			95.0	94.6	1.7	1.9
FPC			**95.5**	**95.2**	**1.5**	**1.6**
Irish nouns						
Lemma	63.5	66.9	56.1	59.6	5.4	5.0
Random Ag.			64.9	68.2	4.2	3.8
FPC			**72.1**	**69.6**	**3.3**	**3.6**

Analysis

Mutual predictability: cell A predicts cell B ↔ cell B predicts cell A

Entropy predictiveness: how easy it is to predict a cell

- I think about this like finding a phoneme; the unpredictable form is usually the UR because although it is the least predictive, it is the most informative

Most of the languages tested (except Arabic) have medium-to-high mutual predictability and high entropy predictiveness: which means that their most unpredictable cells aren't all that useful

Arabic errors: sound vs. broken plurals

Latin errors: multisource predictors accidentally overwrite phonological clues