Zifei Shan

Using Knowledge to Improve Optical Character Recognition

Motivation

- Our system PaleoDeepDive use OCRs to process millions of scanned papers with OCR.
- OCRs have bad quality on these papers.
  - We improved them by Paleo-specific rules.
- We ask:
  - What kinds of knowledge can we use to improve domain-specific OCRs?
  - Can these knowledge be adopted to improve other computer-based tasks?

Knowledge Taxonomy

- We try to use **knowledge** to improve OCR:
  - Dictionary & corpus statistics
  - OCR-specific rules
  - Deep linguistic features (POS, NER, dependency..)
  - Knowledge base: entity linking, relation matching

Initial Error Analysis

- Tool: popular OCRs: *Tesseract* and *Cuneiform*
- Method: Hand-label OCR outputs
- Dataset: 3 documents from PaleoDeepDive
- Result:
  - Average precision: 81% and 73%.
    - 63% and 45% on "dirty" document
  - 7.69% words both OCRs fail

Initial Error Analysis: Takeaways

- We also labeled: what knowledge can solve each error
- Most errors are automatically solvable!
  - An ideal system with all these knowledge can **fix 93.52% errors!**
  - Most useful knowledge (with lesion):
    - Corpus statistics (1-gram of words) 34.7%
    - OCR specific knowledge 17.1%
    - Rules to generate candidates 15.5%
    - Knowledge base (entity linking) 2.6% (avg), 5.2% (dirty)

Examples: Corpus Statistics

Correct	Incorrect	Knowledge
CO2	COz	1-gram
more than	more thin	2-gram
1907 and 1908	%07 and I908	3-gram
No.	No;	1-gram

Examples: OCR-specific Rules

Correct	Incorrect	Knowledge
ENTOPROCTA	ENToPRocTA	case-coherency
144	l44	char-coherency
, limiting this group	. limiting this group	dot -> upper case
first	?rst	"fi -> ?" for Tesseract

Examples: Linguistic Features

Correct	Incorrect	Knowledge
(incomplete)	{incomplete)	dependency path
paper ***is*** based [***VBZ***]	paper ***18*** based [***VBZ***]	word-POS coherency
1982 [DATE]	lpsz [DATE]	word-NER coherency
S. Bur. [PERSON]	S. Bur, [PERSON]	person-dot

Examples: Knowledge base

Correct	Incorrect	Knowledge
Plesiechinus Itawkinsi Jesionek-Szymanska	Plesiechinus hawkinsi Jesionek-Szyma?ska	Entity linking (Freebase)
CALIFORNIA ACADEMY OF SCIENCES	CALIFORNIA ACADEMIA' OF SCIENCES	Entity linking (Freebase)
HETTANGIAN SINEMURIAN TOARCIAN	HETTANGIAN SINEMURIALAI TOARCIAN	Entity linking (PaleoDB Taxonomy)
Upper Miocene; Oeningen	Upper Miocene; Deningen	Relation matching

Examples: Generate Candidates

Existing Candidates	New candidates	Knowledge
Palaeontolngy, pslaeontolosy	Paleontology	Edit distance + corpus statistics
P. echimzta, P. ectzinata	P. echinata	Edit distance + knowledge base
has ulrezidy been, hila slreadv been	has already been	Edit distance + corpus statistics
identi?cation	identification	OCR-specific edit rules
Americo. A m e rican	American	combination
ofthese	of these	segmentation

Putting it together

- We are building an ensemble OCR system using DeepDive.
- Baseline: Ensemble of Tesseract & Cuneiform
- Improve: all useful knowledge above
  - corpus statistics: Google Ngram
  - Deep linguistic features: Stanford NLP parser
  - Knowledge base: Freebase, PaleoDeepDive
  - Distant supervision: 3-gram? *(open question)*

Ongoing System Statistics

- TODO: demonstrate numbers of ongoing system on DD…

Speech recognition?

- Adopt the lessons to speech recognition?
  - Corpus statistics
  - Linguistic features
  - Knowledge base
  - Generate candidates
    - Visual-based -> Audio based edit distance
- Distant Supervision?
  - Labeled corpus? Web? …