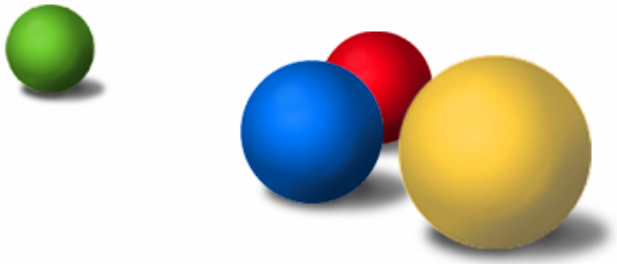
Google

Tesseract OCR Engine

What it is, where it came from, where it is going.

Ray Smith, Google Inc

OSCON 2007



Contents

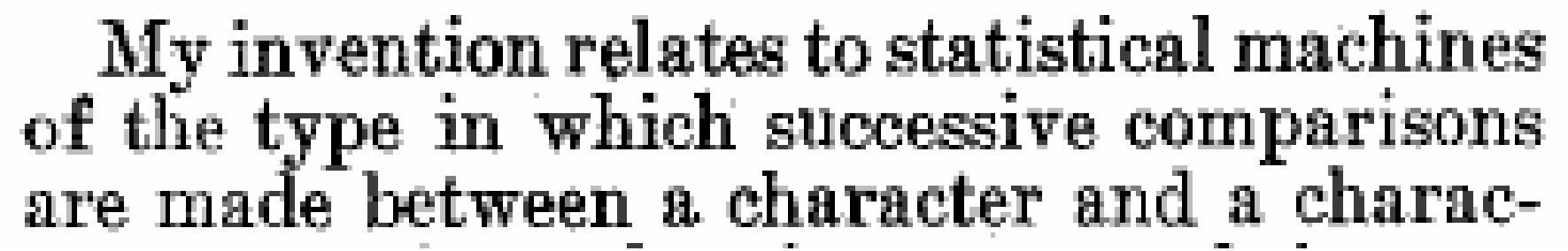
Google

- Introduction & history of OCR
- Tesseract architecture & methods
- Announcing Tesseract 2.00
- Training Tesseract
- Future enhancements

A Brief History of OCR

Google

- What is Optical Character Recognition?


My invention relates to statistical machines of the type in which successive comparisons are made between a character and a charac-

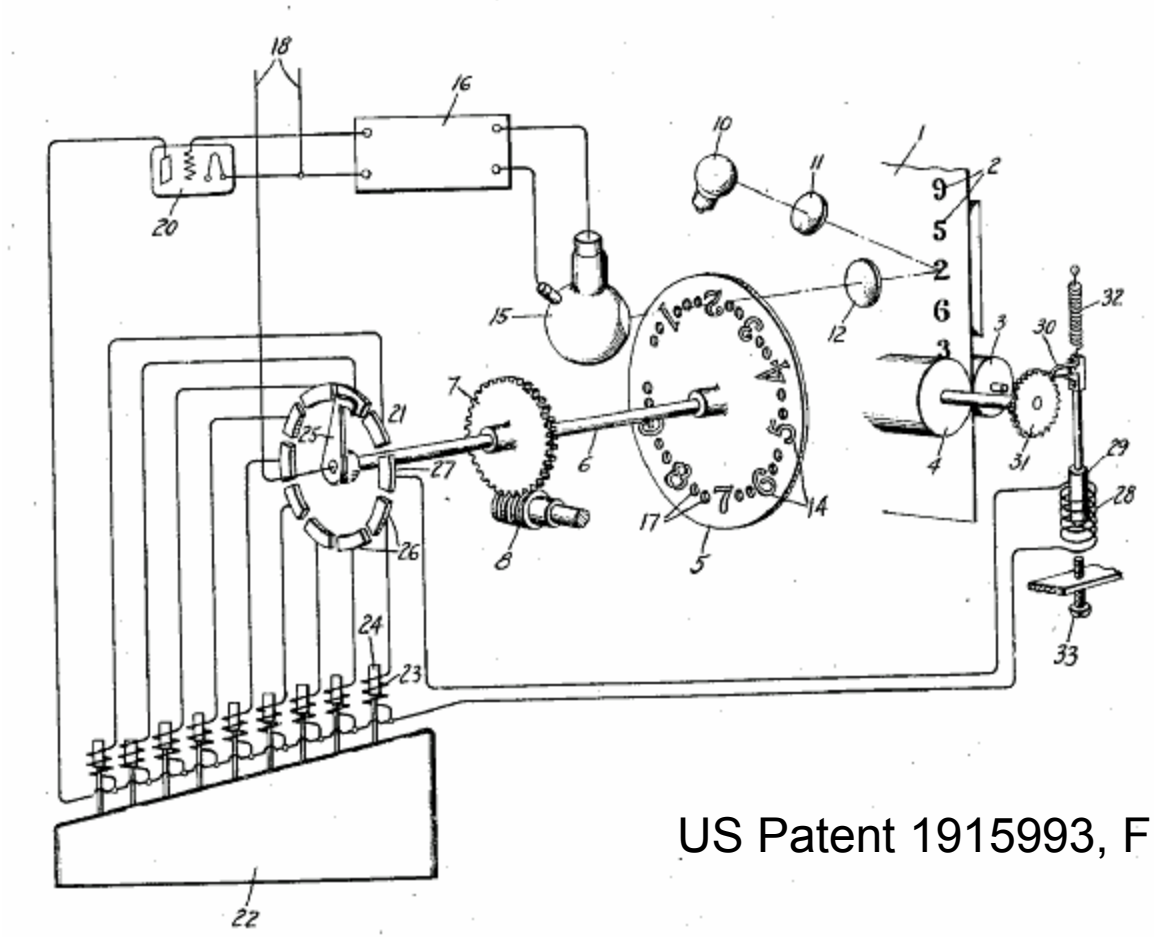
OCR

My invention relates to statistical machines
of the type in which successive comparisons
are made between a character and a charac-

A Brief History of OCR

Google

- OCR predates electronic computers!



US Patent 1915993, Filed Apr 27, 1931

A Brief History of OCR

Google

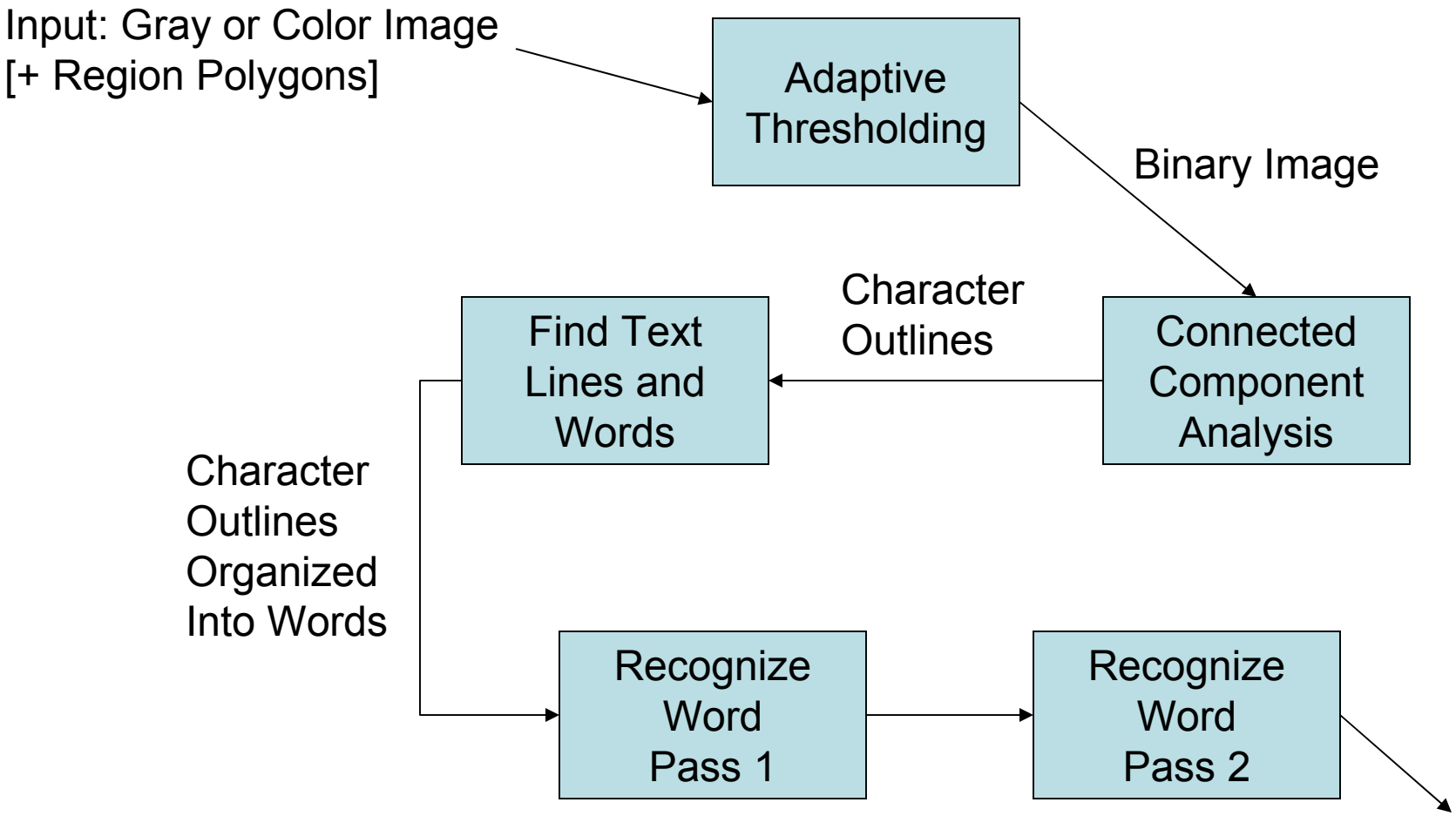
- 1929 – Digit recognition machine
- 1953 – Alphanumeric recognition machine
- 1965 – US Mail sorting
- 1965 – British banking system
- 1976 – Kurzweil reading machine
- 1985 – Hardware-assisted PC software
- 1988 – Software-only PC software
- 1994-2000 – Industry consolidation

Tesseract Background

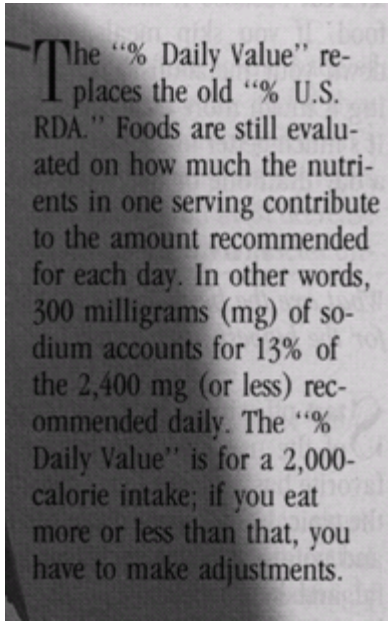
Google

- Developed on HP-UX at HP between 1985 and 1994 to run in a desktop scanner.
- Came neck and neck with Caere and XIS in the 1995 UNLV test.
  (See http://www.isri.unlv.edu/downloads/AT-1995.pdf)
- Never used in an HP product.
- Open sourced in 2005. Now on: http://code.google.com/p/tesseract-ocr
- Highly portable.

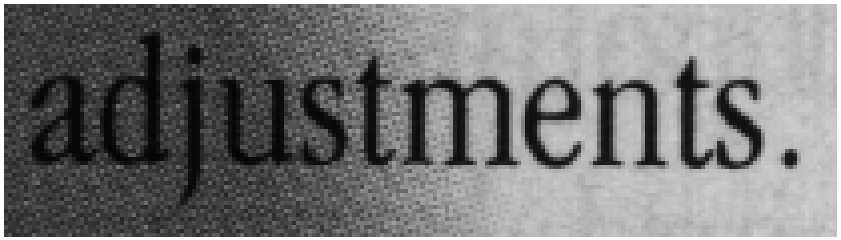
Tesseract OCR Architecture

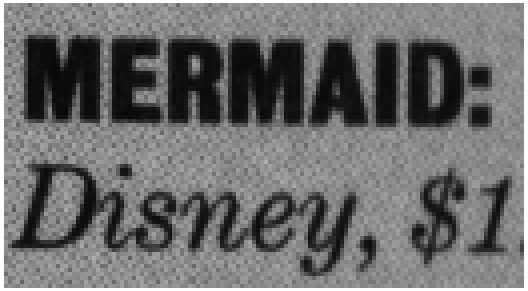


Adaptive Thresholding is Essential


The "% Daily Value" replaces the old "% U.S. RDA." Foods are still evaluated on how much the nutrients in one serving contribute to the amount recommended for each day. In other words, 300 milligrams (mg) of sodium accounts for 13% of the 2,400 mg (or less) recommended daily. The "% Daily Value" is for a 2,000-calorie intake; if you eat more or less than that, you have to make adjustments.

Some examples of how difficult it can be to make a binary image
Taken from the UNLV Magazine set.
(http://www.isri.unlv.edu/ISRI/OCRtk)


adjustments.


MERMAID:
Disney, $1


'Secret Games 2: The Escort'
'Chained Heat 2'
'Amityville: A New Generation'
'Lady Dragon 2'
'Relentless 3'
What 4?
UNATTRIBUTED

VINTAGE

DAVID HOLZMAN'S DIARY L.M. Kit Carson *(1967, Fox Lorber, unrated, $79.95)* Years before he went Hollywood with *The Big Easy* and *Great Balls of Fire!*, director Jim McBride scored with this droll put-on, a fake underground movie that shoots snarky darts through film-student pretensions. David Holzman (Carson) is a sincere young twerp who, seeking Big Truth, films his life in what was the then-fashionable grainy B&W vérité. What he finds is a swift corrective to all those who think movies are more important than reality. Raw, ragged, and right. *Diary* is both an

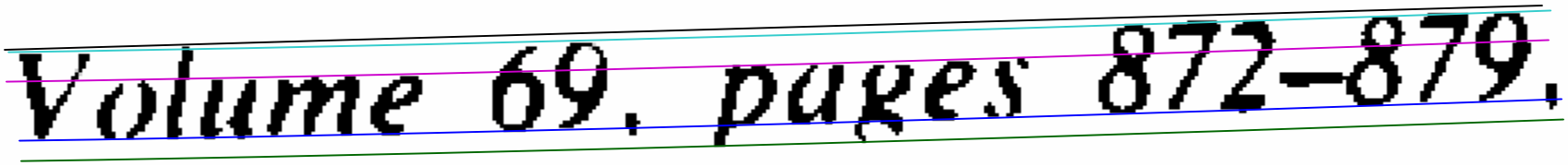
'CRYING' TIME

WITH ARTSIER video renters having run through *Howards End* (which drops off the top 10 this week), *The Crying Game* finally joins the chart at No. 7 in its third week in release. Apparently, laserdisc buyers were awaiting the Oscar nominee far more eagerly than tape renters: The laser version debuts at a surprising No. 2.

Meanwhile, the upward mobility of Columbia TriStar/Voyager's laserdisc of *Bram Stoker's Dracula* may soon be halted. Two just-released inexpensive, wide-screen versions should cut into the sales of the current pricey

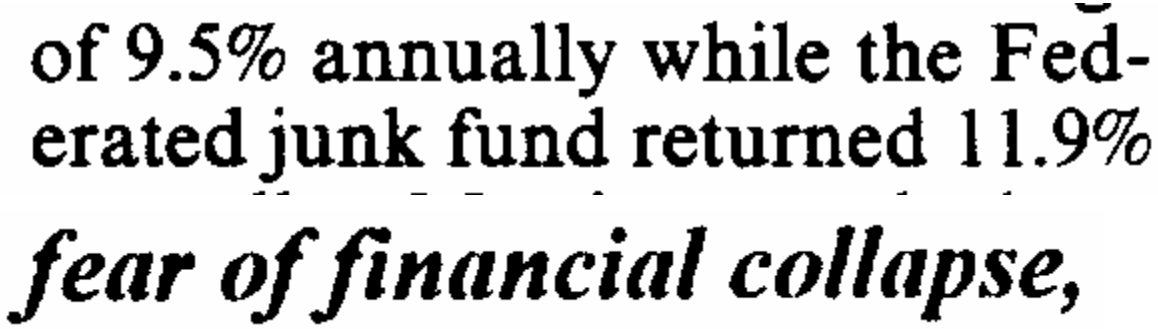
Baselines are rarely perfectly straight

- Text Line Finding – skew independent – published at ICDAR'95 Montreal. (http://scholar.google.com/scholar?q=skew+detection+smith)
- Baselines are approximated by quadratic splines to account for skew and curl.
- Meanline, ascender and descender lines are a constant displacement from baseline.
- Critical value is the x-height.

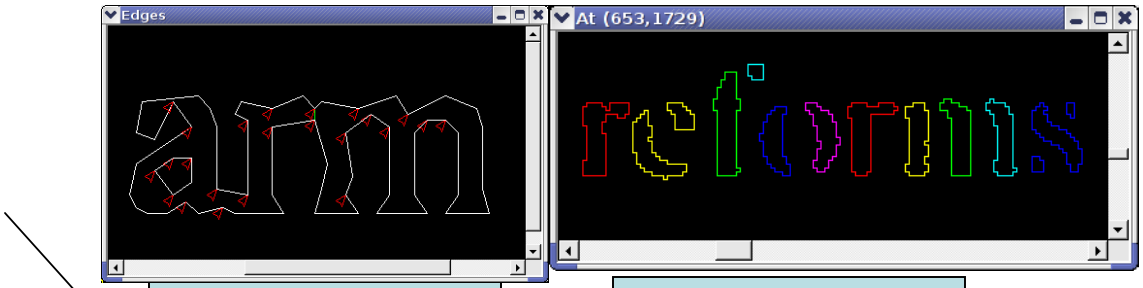


Spaces between words are tricky too Google

- Italics, digits, punctuation all create special-case font-dependent spacing.
- Fully justified text in narrow columns can have vastly varying spacing on different lines.



Tesseract: Recognize Word



Character Chopper

Character Associator

Done?

No

Yes

Adapt to Word

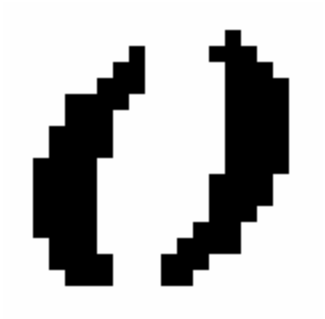
Static Character Classifier

Dictionary

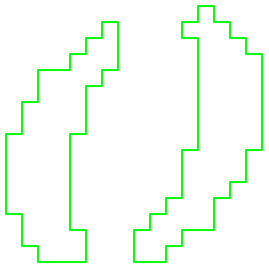
Adaptive Character Classifier

Number Parser

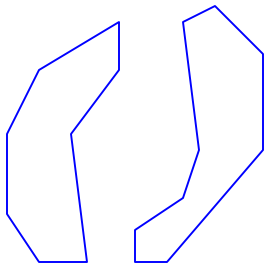
Outline Approximation



Original Image



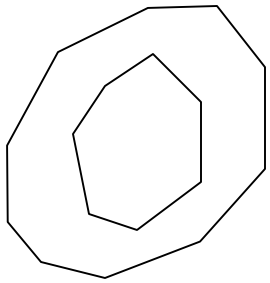
Outlines of components



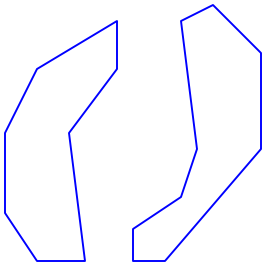
Polygonal Approximation

Polygonal approximation is a double-edged sword.
Noise and some pertinent information are both lost.

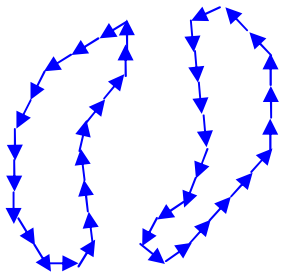
Tesseract: Features and Matching



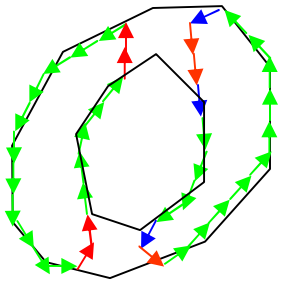
Prototype



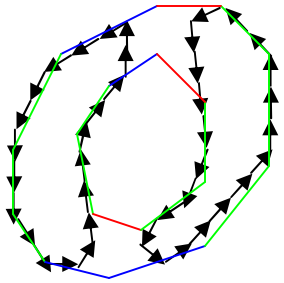
Character to classify



Extracted Features



Match of Prototype To Features



Match of Features To Prototype

- Static classifier uses outline fragments as features. Broken characters are easily recognizable by a small->large matching process in classifier. (This is slow.)
- Adaptive classifier uses the same technique! (Apart from normalization method.)

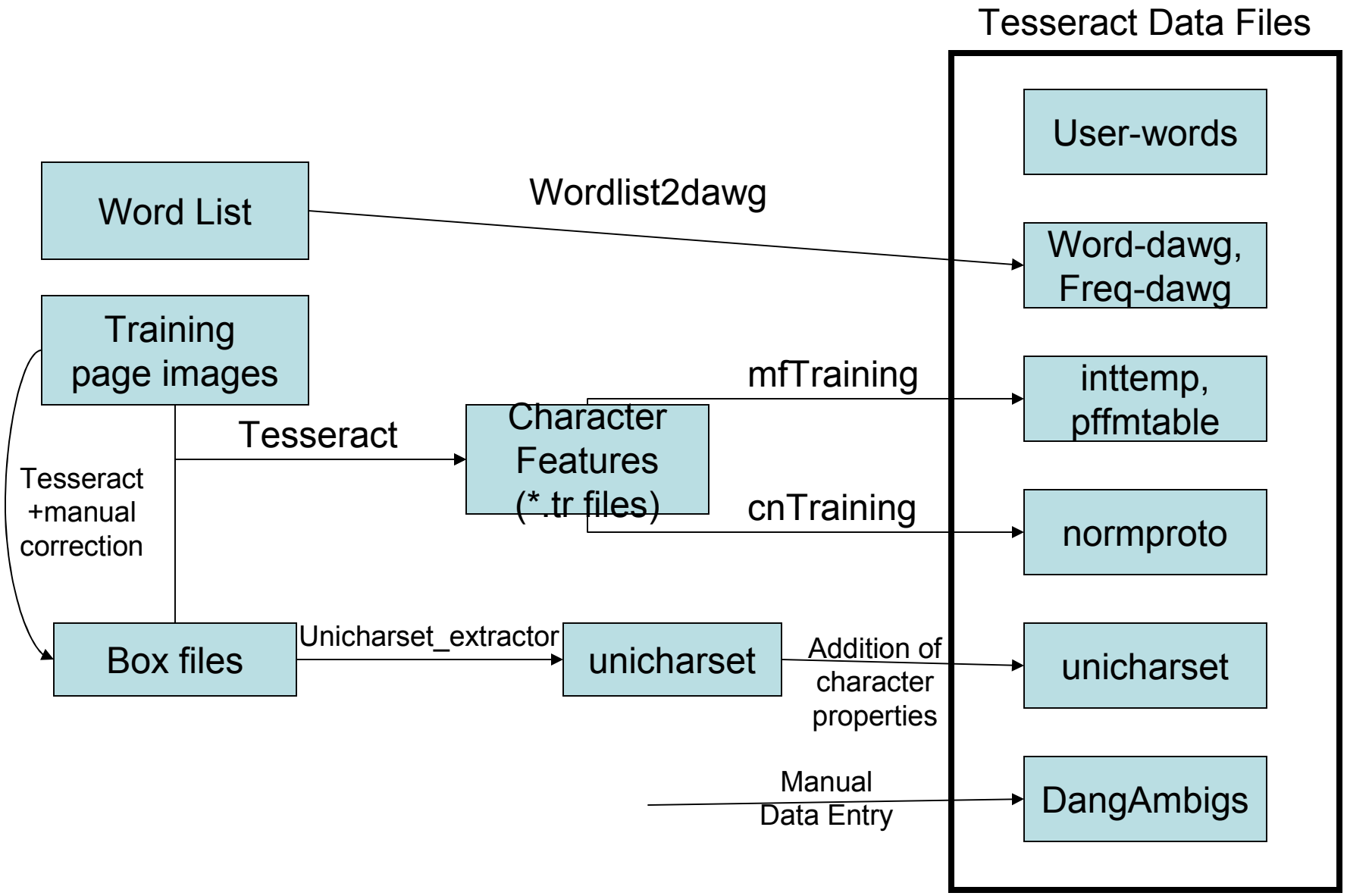
Announcing tesseract-2.00

Google

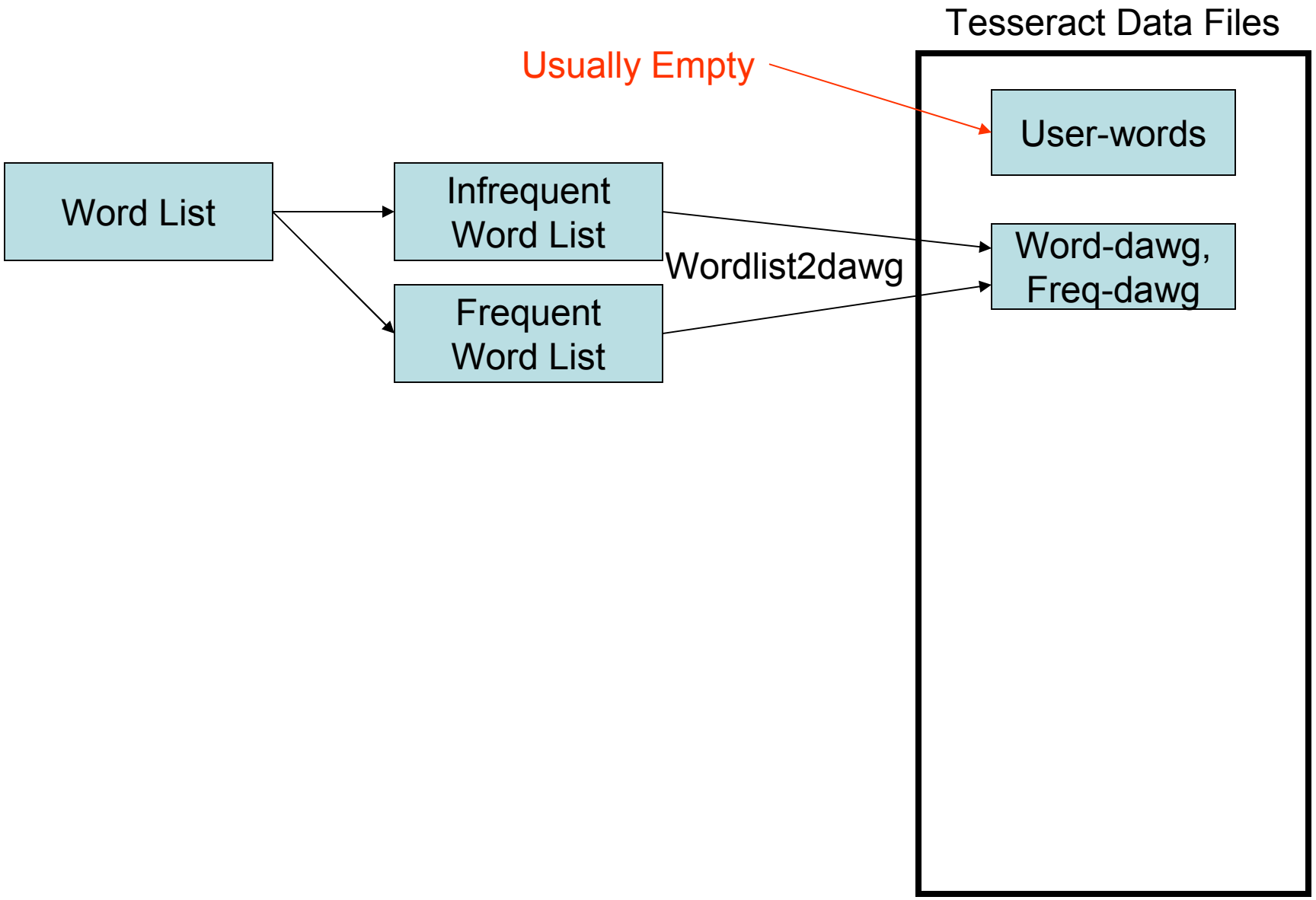
- Fully Unicode (UTF-8) capable
- Already trained for 6 Latin-based languages (Eng, Fra, Ita, Deu, Spa, Nld)
- Code and documented process to train at http://code.google.com/p/tesseract-ocr
- UNLV regression test framework
- Other minor fixes

Training Tesseract

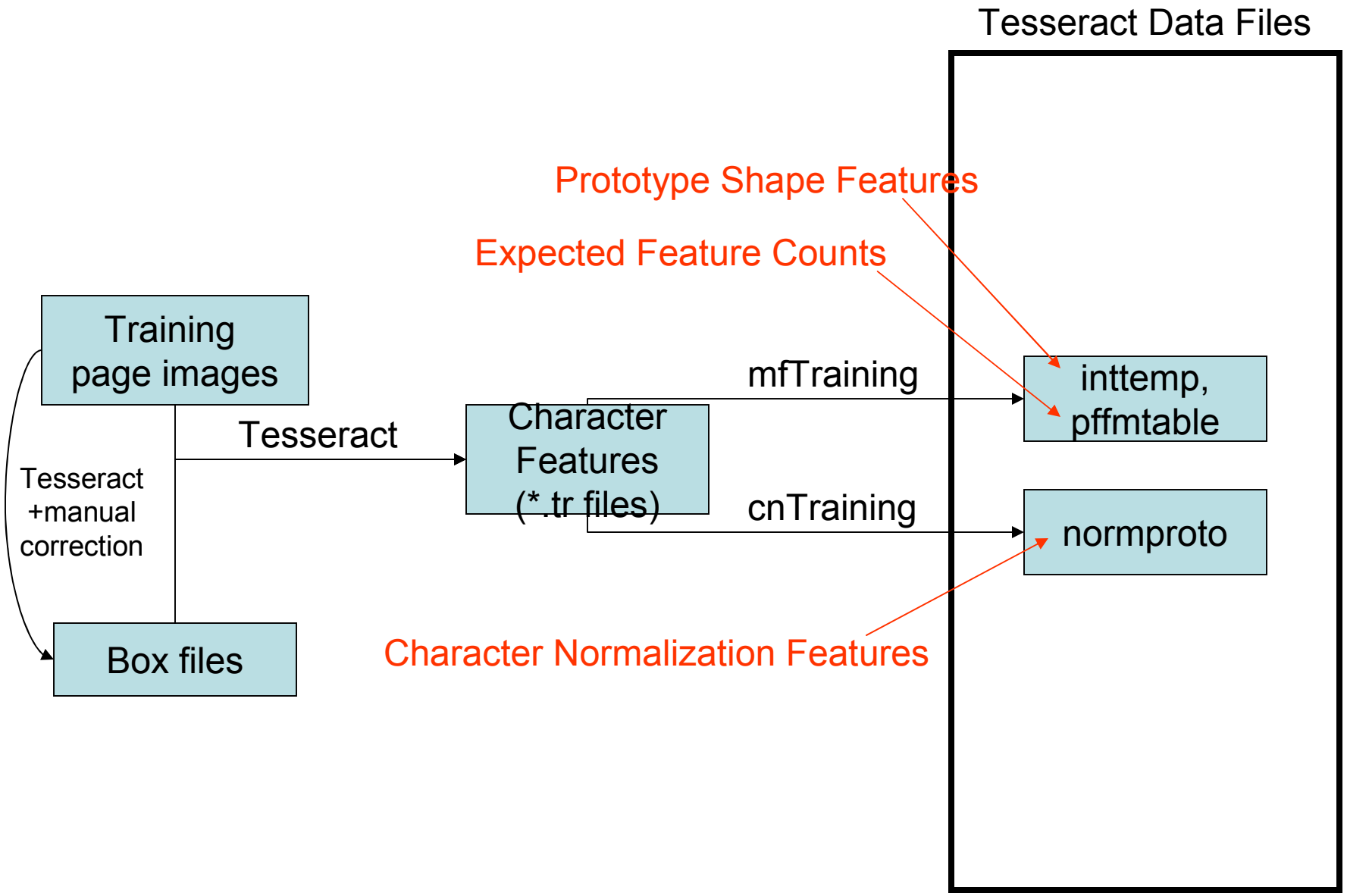
Google



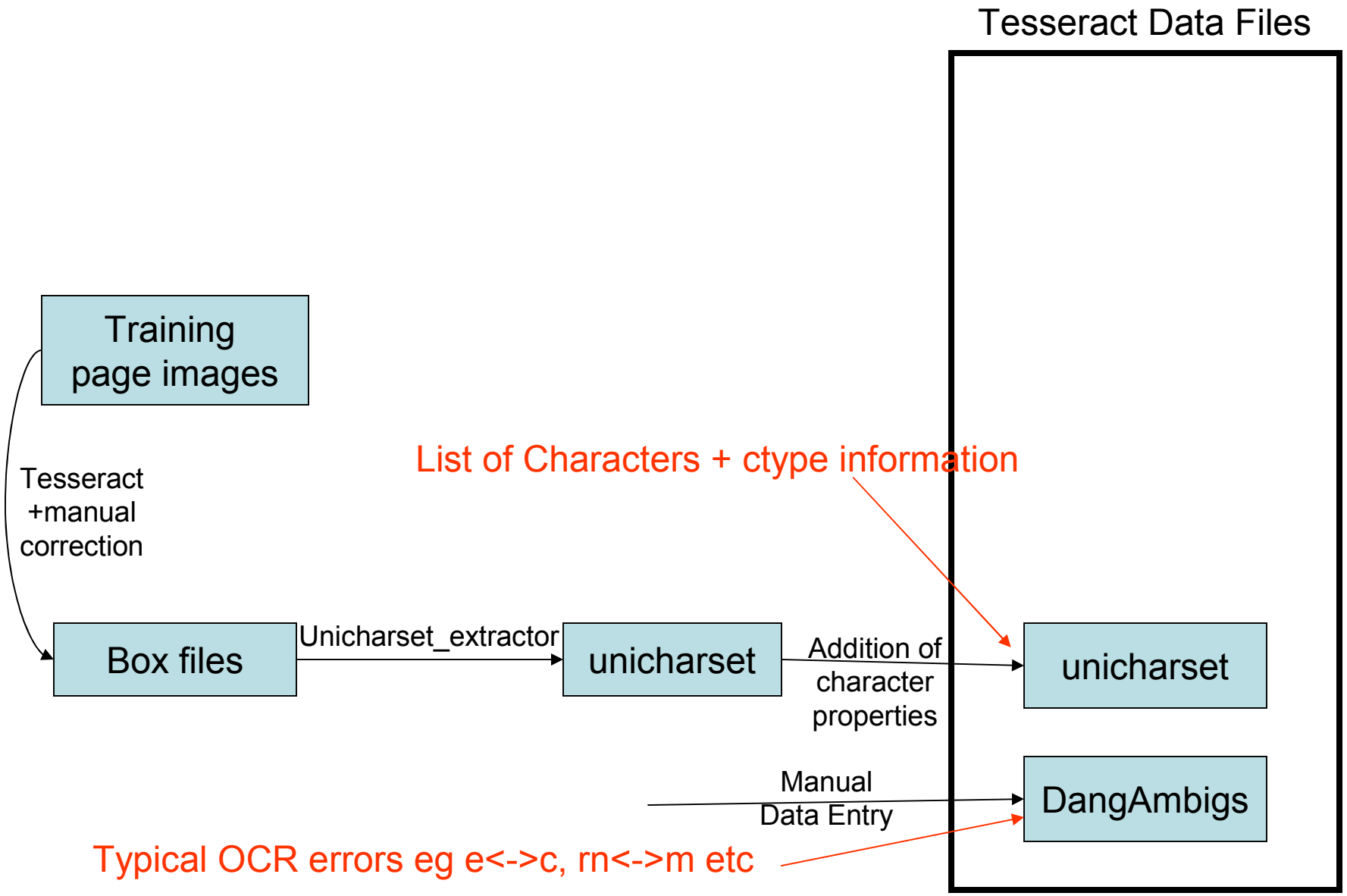
Tesseract Dictionaries



Tesseract Shape Data



Tesseract Character Data



Accuracy Results

Google

Comparison of current results against 1995 UNLV results

Testid	Testset	Character			Non-stopword		
		Errors	Accuracy	Change	Errors	Accuracy	Change
1995	Bus.3B	5959	98.14%		1293	95.73%	
1995	Doe3.3B	36349	97.52%		7042	94.87%	
1995	Mag.3B	15043	97.74%		3379	94.99%	
1995	News.3B	6432	98.69%		1502	96.94%	
Gcc4.1	Bus.3B	6258	98.04%	5.02%	1312	95.67%	1.47%
Gcc4.1	Doe3.3B	28589	98.05%	-21.35%	6692	95.12%	-4.97%
Gcc4.1	Mag.3B	14800	97.78%	-1.62%	3123	95.37%	-7.58%
Gcc4.1	News.3B	7524	98.47%	16.98%	1220	97.51%	-18.77%
Gcc4.1	Total	57171		-10.37%	12347		-6.58%

Commercial OCR v Tesseract

Google

- 100+ languages.
- Accuracy is good now.
- Sophisticated app with complex UI.
- Works on complex magazine pages.
- Windows Mostly.
- Costs $130-$500

- 6 languages + growing.
- Accuracy was good in 1995.
- No UI yet.
- Page layout analysis coming soon.
- Runs on Linux, Mac, Windows, more...
- Open source – Free!

Tesseract Future

Google

- Page layout analysis.
- More languages.
- Improve accuracy.
- Add a UI.

The End

Google

- For more information see: http://code.google.com/p/tesseract-ocr