COMPARING A LINGUISTIC AND A STOCHASTIC TAGGER

Which is which

- The paper mentions a statistical and a stochastic tagger
- Should be a statistical and a linguistic tagger
- Terms are used interchangeably in the paper
- Stochastic means something relating to conjecture or randomness

What was compared

- The performance of a tagger based on a set of linguistic constraints
- The performance of a tagger based on statistical analysis of text

The linguistic tagger

- An updated version of the EngCG tagger – EngCG2
- 3600 hand coded constraints!
- Consists of the following sequentially applied modules:
  - Tokenization
  - A morphological analysis consisting of:
    - A Lexical component
    - A rule based guesser for unknown words
  - Resolution of morphological ambiguities

The linguistic tagger cont.

- Operates on a reduced tagset of its own
- Tagset is grammatically rather than semantically motivated
- Would most likely need to be rewritten to a large degree for other languages
- Highly accurate with low ambiguity

The statistical tagger

- A classical trigram-based Hidden Markov Models decoder
- Calculates the most likely tag for a word based on the words surrounding it
- Creates a reverse suffix tree
- Statistical smoothing is performed by reversing up the tree

The statistical tagger cont.

- Employs a more widely used tagset
- Can be used on any language provided a sufficiently large corpus exists
- Learning curve levels off at around 322.000 words for english
- Acquires increased accuracy at the cost of increased ambiguity

The purpose

- To quell criticism of the EngCG tagger
- Such as:
  - The tagset is overly simplistic
  - The accuracy of the EngCG tagger has been overstated
  - EngCG trades off high accuracy for high ambiguity

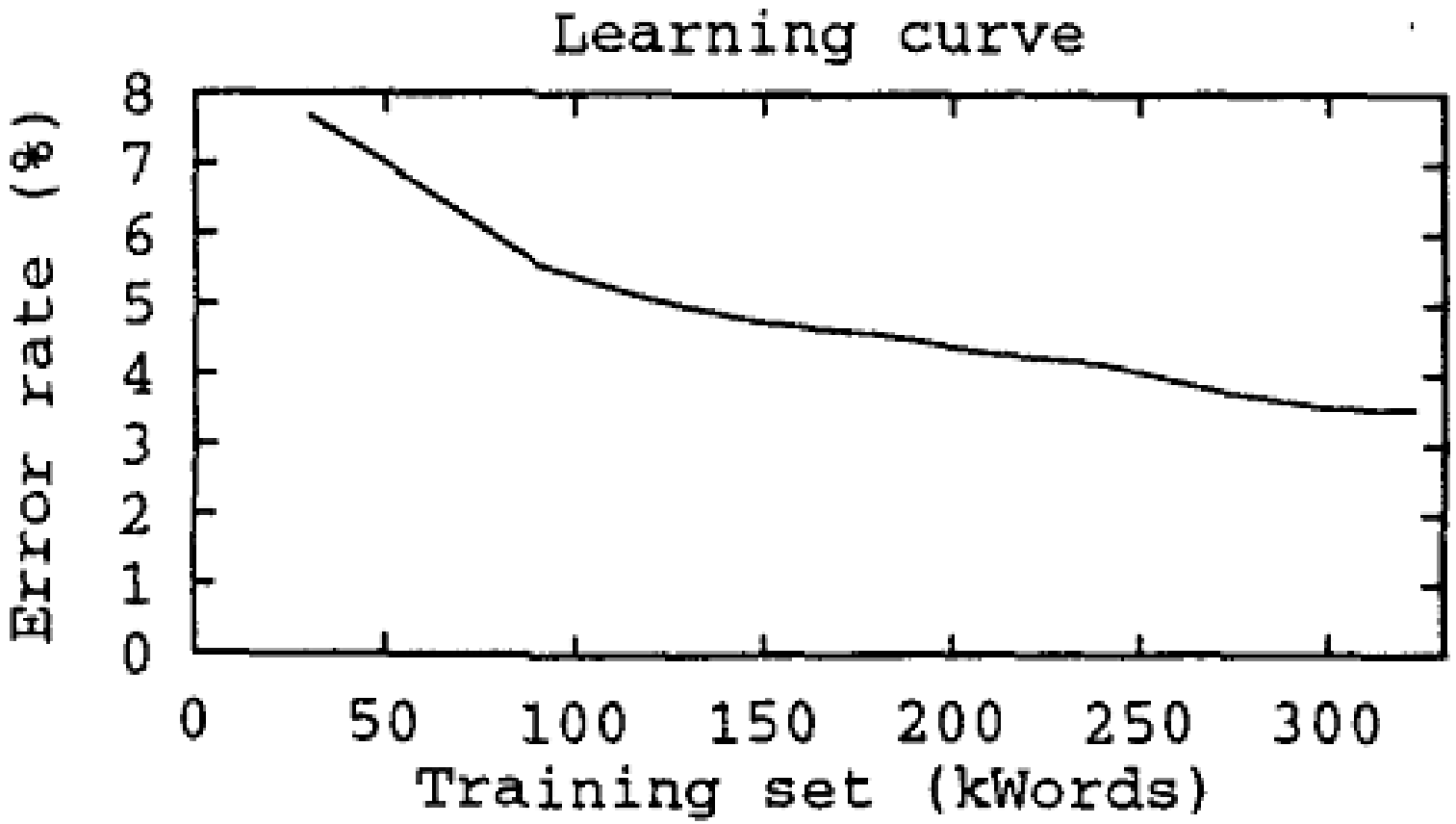
The setup

- A sample of 357.000 words from the Brown corpus
- Tagged by EngCG and human corrected where needed
- A held out benchmark corpus of ~55.000 words from various texts
- Annotated by preprocessor and morphological analyser.
- Fully disambiguated by 2 experts

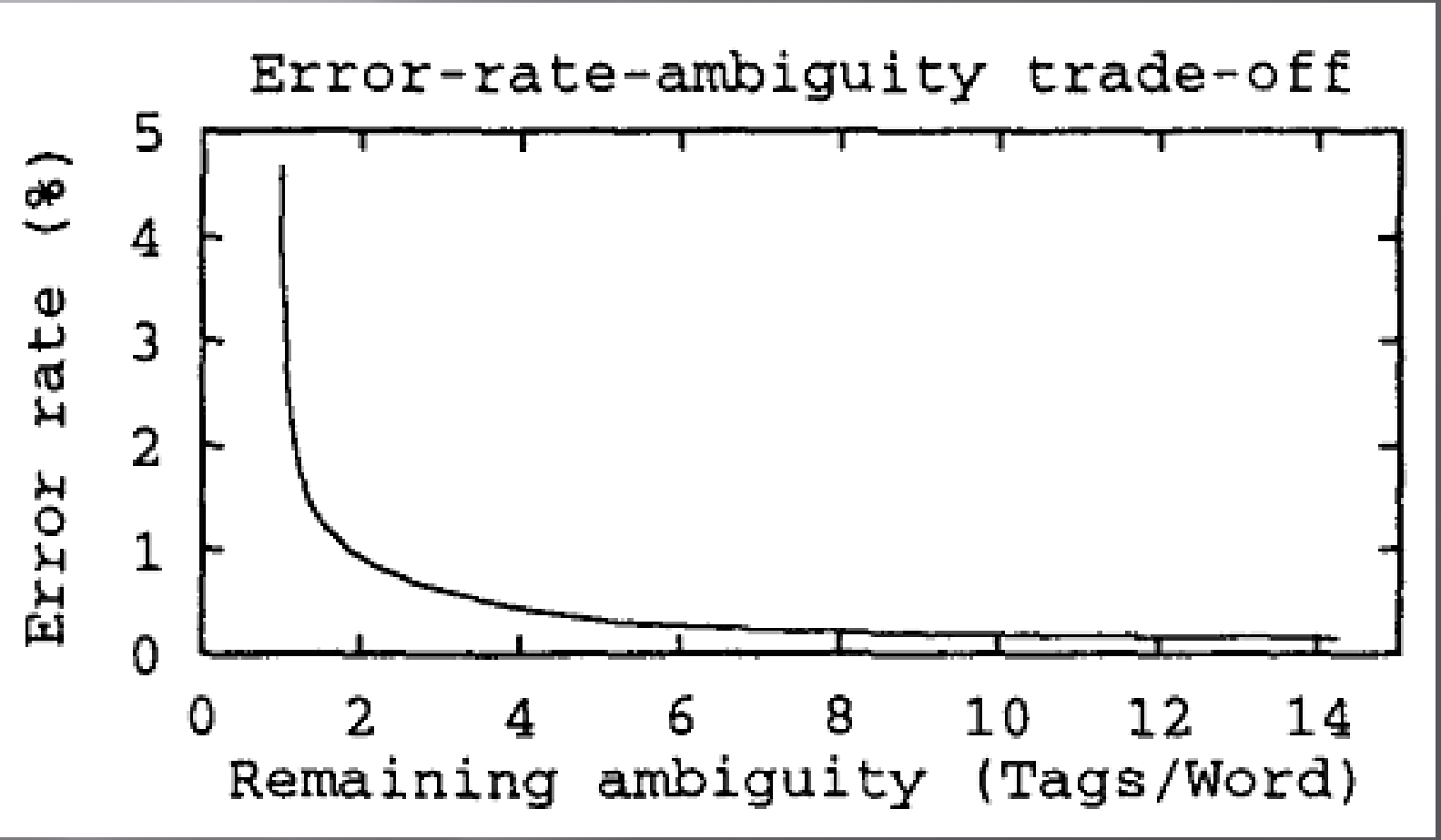
The testing

- Each tagger was run on the benchmark corpus
- Error rates and ambiguity compared
- Also the learning curve for the statistical tagger was measured

Statistical tagger's learning curve



Statistical analyser's amiguity/accuracy tradeoff



The results

- EngCG2 performed radically better
- The statistical tagger's error rates higher by a factor of 8.6 to 28.0!
- The ambiguity remained very low for very low error rates: 0.10% at 1.070 tags per word

The results cont.

Ambiguity (Tags/word)	Error rate (%) Statistical Tagger (δ)	(γ)	EngCG
1.000	4.72	4.68	
1.012		4.20	
1.025		3.75	
1.026		(3.72)	0.43
1.035		(3.48)	0.29
1.038		3.40	
1.048		(3.20)	0.15
1.051		3.14	
1.059		(2.99)	0.12
1.065		2.87	
1.070		(2.80)	0.10
1.078		2.69	
1.093		2.55	

Caveats

- The new version of the EngCG tagger had been created using the Brown corpus as a benchmark
- The EngCG2 tagger had therefore been trained on the benchmark corpus
- The radically better performance required 3600 hand coded rules!