

Entity Summaries

Jing Jiang and Xu Lin
BeeSpace Programmers' Meeting
Sept. 6, 2006

A quick review of the NER component

- Use two types of information to make a prediction
 - Word features and word surface features
 - E.g. p53, XXXless
 - Contextual features
 - E.g. XXX expression, XXX mutants
- Prediction of the same word/phrase is context-sensitive

Examples of Some Ambiguous Gene Names

- **foraging**

- We assayed response decrement for natural and mutant rover and sitter alleles of the foraging (for) gene that encodes a Drosophila PKG. (FN)
- Hybrid disadvantage in the larval foraging behaviour of the two neotropical species of Drosophila pavani and Drosophila gaucha... (TN)

Examples of Some Ambiguous Gene Names

- **SS**

- ...SmZF1 binds both ds and ss DNA oligonucleotides,... (TN)
- Coexpression of Ss and Tgo in Drosophila SL2 cells... (TP)
- The origin of germline-limited chromosomes (Ks) as descendants of somatic chromosomes (Ss) and their... (FP)

Examples of Some Ambiguous Gene Names

- **black**

- The purpose of this study was to investigate the black gene, and protein,... (FN)
- ...beta-alanine biosynthesis is regulated by black. (FN)
- Screening a cDNA library prepared from silk-producing glands of the black widow spider,... (TN)

Examples of Some Ambiguous Gene Names

- **clock**

- ...a novel fitness-related phenotype may be linked to noncircadian expression of clock genes in the ovaries. (TP)
- ...mPer1 could operate in the adaptation of the circadian clock of nocturnal mice to... (TN)

Examples of Some Ambiguous Gene Names

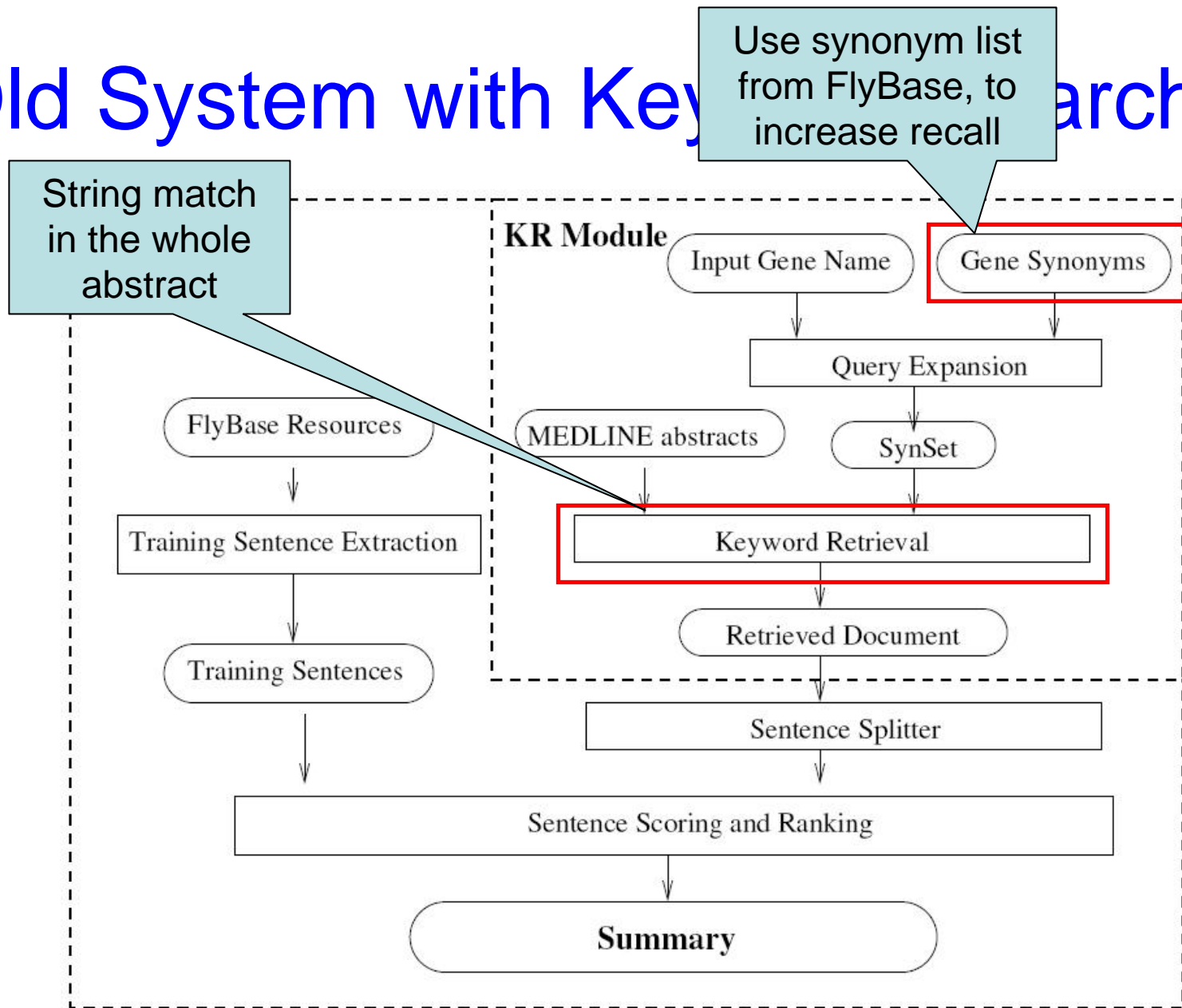
- **ERG**
 - To establish the predicted existence of a *Drosophila* gene in the erg subfamily and... (FN)
 - The ERG analysis of the *norpA* mutants suggests that... (TN)
 - Here we show that the electroretinogram (ERG), the extracellular recording...(FP)

Examples of Some Ambiguous Gene Names

- **pdf**

- PDF is coded in a precursor protein together with another neuropeptide... (TP)
- ...the Drosophila brain that express the period (per) and pigment dispersing factor (pdf) genes play... (TP)

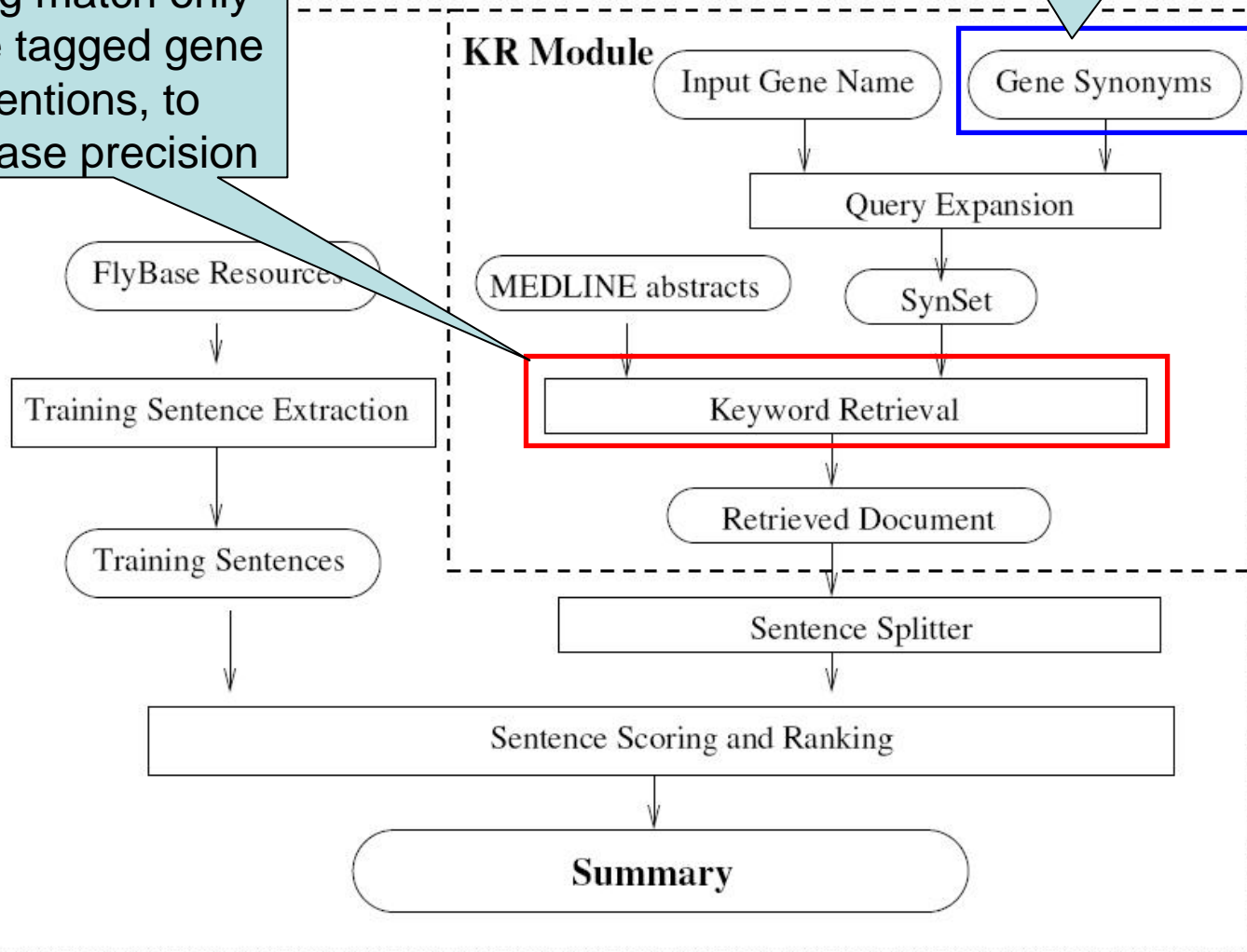
Old System with Key Search



New System w

Will be replaced by an automated normalizer if there is

String match only in the tagged gene mentions, to increase precision



Changes with integration of NER

- Recall
 - Tokenization
 - Keyword match (whole abstract => gene mentions)
 - Synonym list
- Precision
 - Exact match => exact match but allowing crossing tag boundary

Example

- Query: ABC-a

```
xxxxxxxxxxxxxxxx<GENE>ABC a</GENE> gene xxxxxxxxxxxxxxxxxxx  
xxxxxxxxxxxxxxxxxxxxxxxx ABC a xxxxxxxxxxxxxxxxxxxxxxxxxxx  
xxxxxxxxxxxxxxxx<GENE>ABC</GENE> a encodes xxxxxxxxxxx  
xxx<GENE> gene ABC a</GENE> xxxxxxxxxxxxxxxxxxx
```

- Without NER: match all 4 cases
- With NER: not match the second case

Effects of NER on Gene Summarizer

- FP → TN (increase precision)
 - ...mPer1 could operate in the adaptation of the circadian clock of nocturnal mice to... (TN)
- TP → FN (decrease recall)
 - ...beta-alanine biosynthesis is regulated by black. (FN)
- FP → FP (no effect, but not what we want)
 - Here we show that the electroretinogram (ERG), the extracellular recording...(FP)

A Different Approach

- Motivation: to solve a simpler problem than NER because we already know the gene name and its synonyms
- Approach: build a classifier that focuses on contextual features to identify FPs
 - Only use contextual features because the term/phrase already matches a gene name
 - Need some “good” negative examples (ambiguous gene names) in the training data