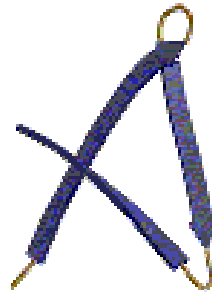




FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA

xldb-Research Group



*Architecture et
Fonction des
Macromolécules
Biologiques*

FiGO: Finding GO Terms in Unstructured Text

Francisco M. Couto,
Mário J. Silva and Pedro Coutinho



Outline

v Introduction

v Method

v Results

v Conclusions



Main Idea

- v Information Content of GO terms' words
- v Example: “pant binding”
 - v The probability of the term being mentioned is higher if “pant” occurs than if only “binding” occurs
 - v Because “binding” is also used in many other terms
 - v “pant” is more informative than “binding”



Information Content

- v Inversely proportional to the number of occurrences
- v Information Content of an object Φ :

$$IC(\Phi) = -\log(\#\Phi/\#\max)$$

- v $\#\Phi$ represents the number of times that Φ occurs
- v $\#\max$ represents the maximum number of times that an object occurs
- v Log function just to improve the calculation



Outline

v Introduction

v Method

v Results

v Conclusions



FiGO

- v Input:
 - v A collection of terms
 - v A piece of text
- v Output:
 - v A ranked list of terms mentioned on the piece of text



Pre-Processing (1)

- v Identify all words present in the terms' names
- v Ignore the stop words
 - v Such as: 'in' or 'on'
- v Compute the number of occurrences of each word
 $\#w$
 - v $\#w$ is the number of terms that have w in its name
- v Compute the information content of each word:

$$IC(w) = -\log(\#w/\#max)$$



Pre-Processing (2)

- For each term's name n composed by the words: w_0, \dots, w_k compute the information content of n :

$$IC(n) = \sum IC(w_i)$$

- Compute the information content of the term's names n_0, \dots, n_k :

$$IC(t) = \max \{ IC(n_i) \}$$



Example

- v Considering:
 - v The term $t = \text{'punt binding'}$
 - v With the synonym 'punt activity'
 - v $\#punt=1, \#binding=4, \#activity=8, \#max=16$
- v $IC(\text{'punt'}) = -\log(1/16) = 4,$
- v $IC(\text{'binding'}) = -\log(4/16) = 2$
- v $IC(\text{'activity'}) = -\log(8/16) = 1$
- v $IC(\text{'punt binding'}) = 4 + 2 = 6$
- v $IC(\text{'punt activity'}) = 4 + 1 = 5$
- v $IC(t) = \max\{6, 5\} = 6$



Procedure (1)

- ▼ Compute the local information content of each term t in the piece of text p :

$$\text{LIC}(t,p) = \sum \text{IC}(w_i)$$

- ▼ Where w_0, \dots, w_1 are words of the term's name that occur in p
- ▼ $\text{IC}(t) \geq \text{LIC}(t,p)$
- ▼ $\text{LIC}(t,p)/\text{IC}(t)$ measures how much of t 's name occurs in p



Procedure (2)

- v Which terms FiGO considers mentioned?
- v Each term t whose:

$$\text{LIC}(t,p) \geq \alpha \times \text{IC}(t)$$

- v $\alpha \in [0,1]$
- v When $\alpha=1$ FiGO selects only the terms fully mentioned
- v When $\alpha=0$ FiGO selects all terms



Example

- v Considering:
 - v The term $t = \text{"punt binding"}$
 - v The pieces of text:
 1. “The protein has a binding activity”
 2. “The protein has a punt activity”
 3. “The protein has a punt binding activity”
 - v $IC(\text{"punt"})=4$ and $IC(\text{"binding"})=2$
- v $IC(t)=6$
- v $LIC(t, p_1)=2$ ($\alpha \leq 1/3$)
- v $LIC(t, p_2)=4$ ($\alpha \leq 2/3$)
- v $LIC(t, p_3)=6$ ($\alpha \leq 1$)



Task 2.1

- v Piece of text = sentence
- v FiGO returned a list of sentences where the term occurred
- v Which sentence?
 - v Containing the protein name
 - v Having the larger LIC
- v If there was no sentence?
 - v The most similar term (FuSSiMeG)



Task 2.2

- v Piece of text = sentence
- v FiGO returned a list of terms with the sentences where they were mentioned
- v Which terms?
 - v The protein name in the same sentence
 - v The most meaningful annotations
 - v The most infrequently annotated terms



Outline

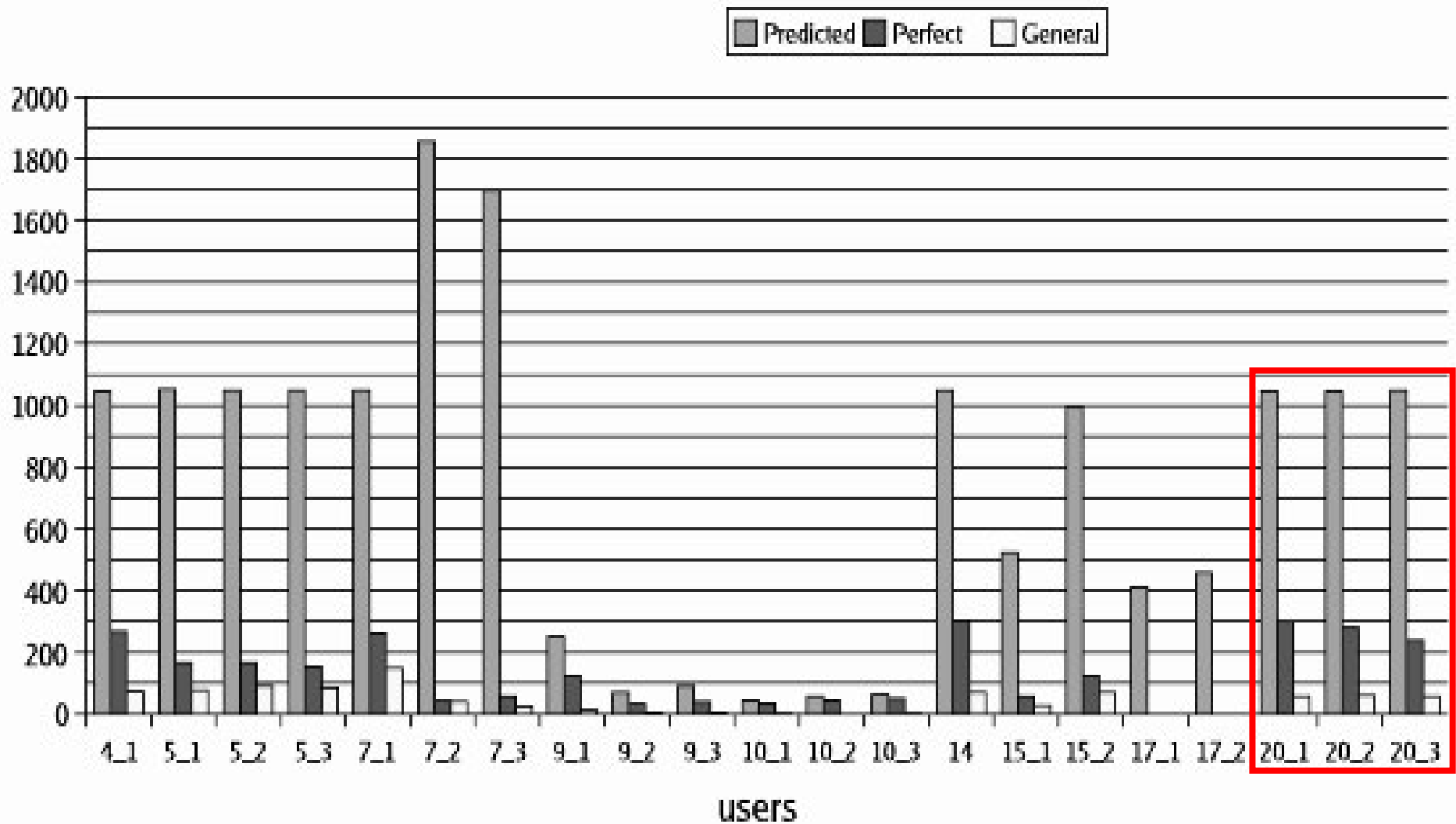
v Introduction

v Method

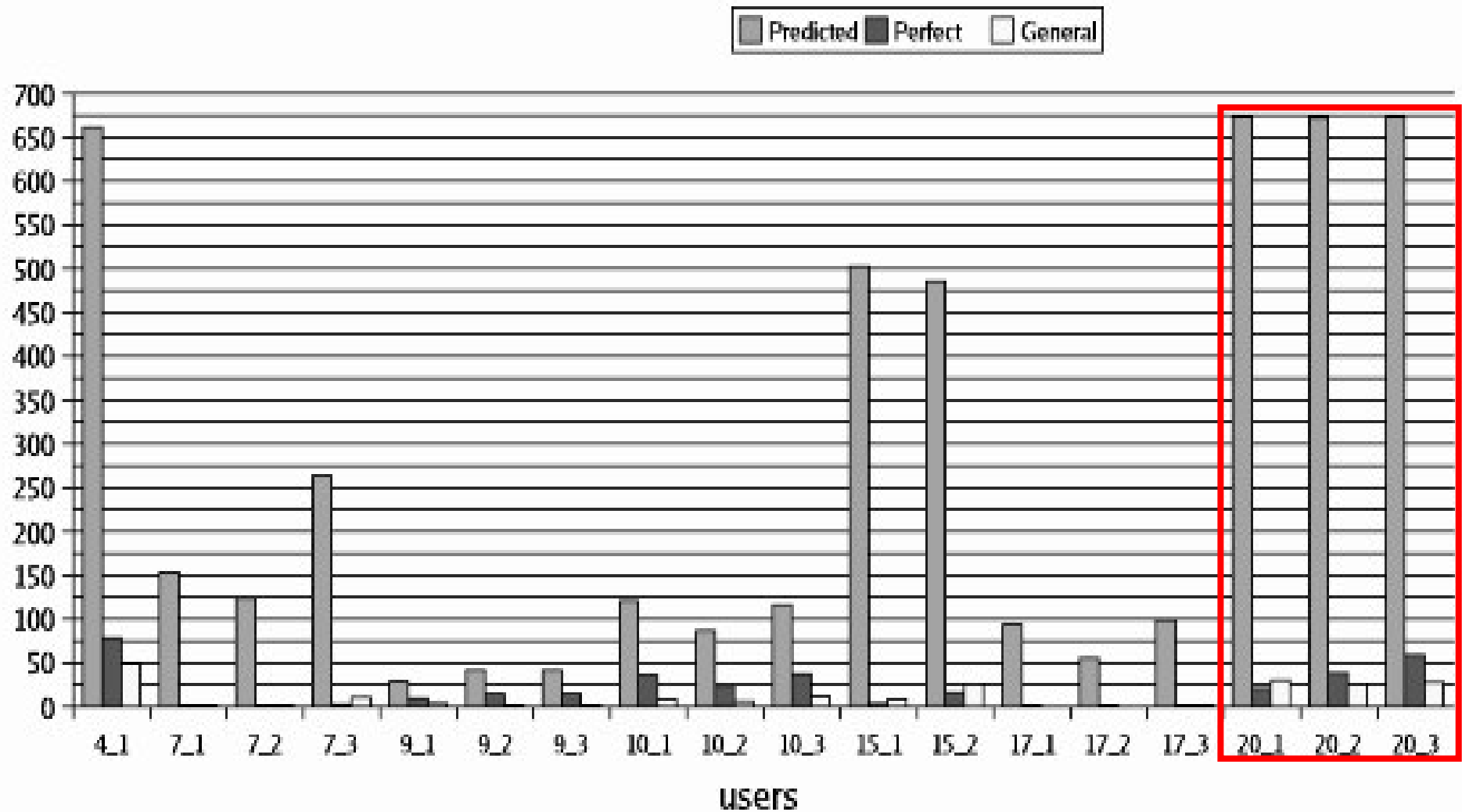
v Results

v Conclusions

Task 2.1 Results



Task 2.2 Results

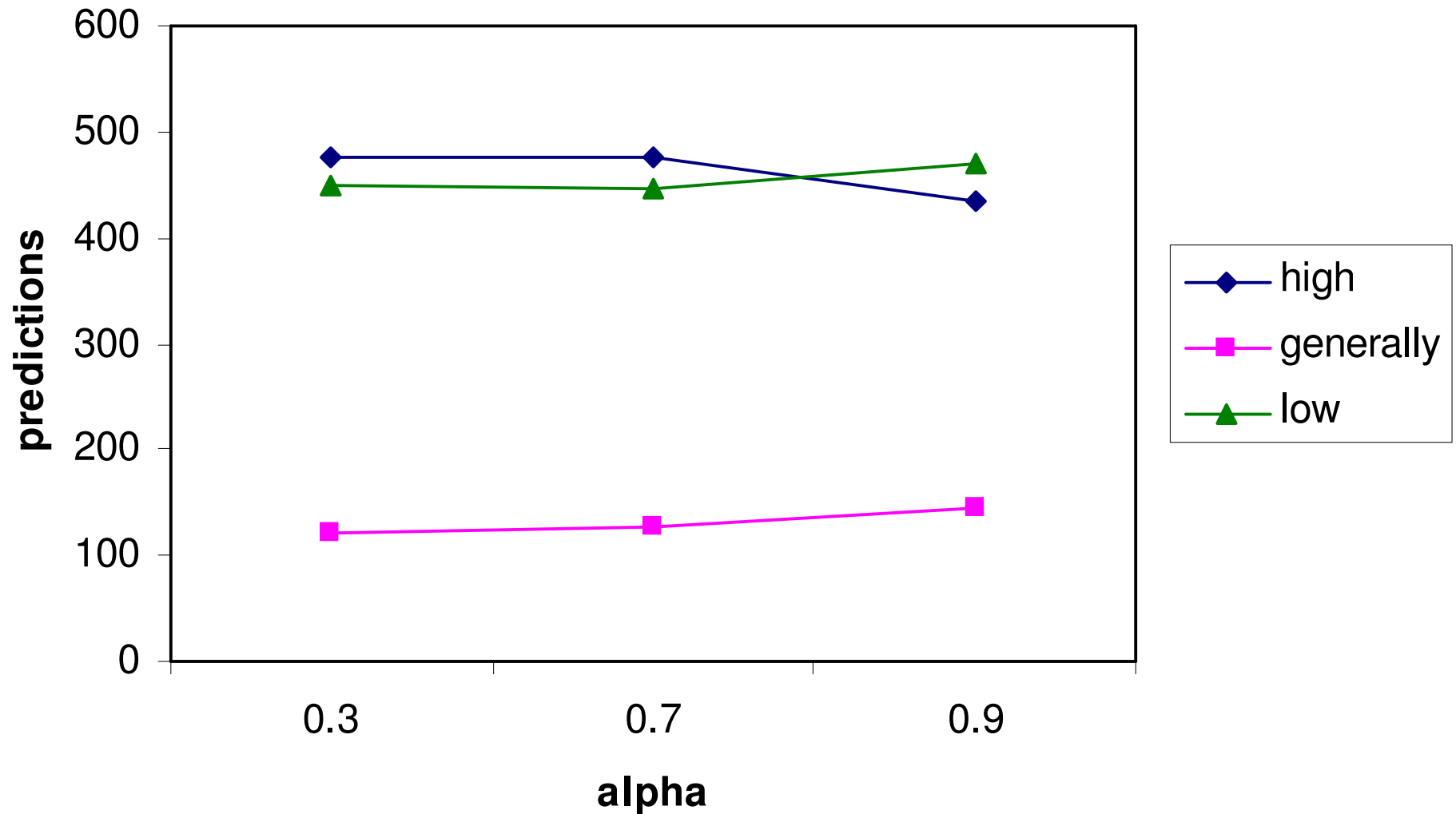




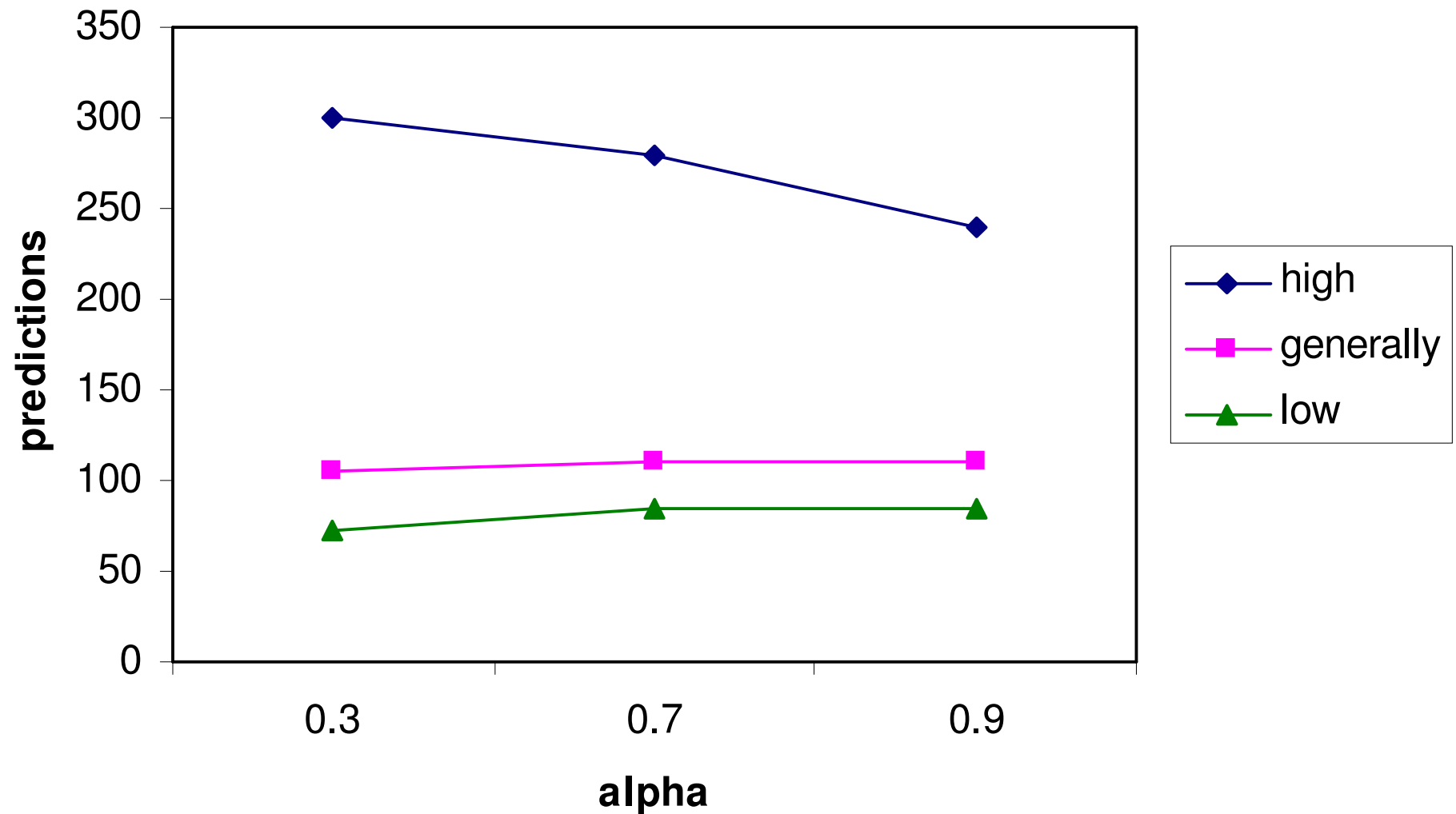
Discussion

- v Very close to the largest number of perfect predictions achieved
- v Our accuracy was very far from the the best results
 - v Submitted the expected number of predictions
 - v Some had a low confidence level
 - v Filter predictions according to their confidence level to achieve better accuracy
- v Better performance in task 2.1 than in task 2.2 due to greater difficulty of task 2.2

Task 2.1 GO Evaluation



Task 2.1 Protein Evaluation

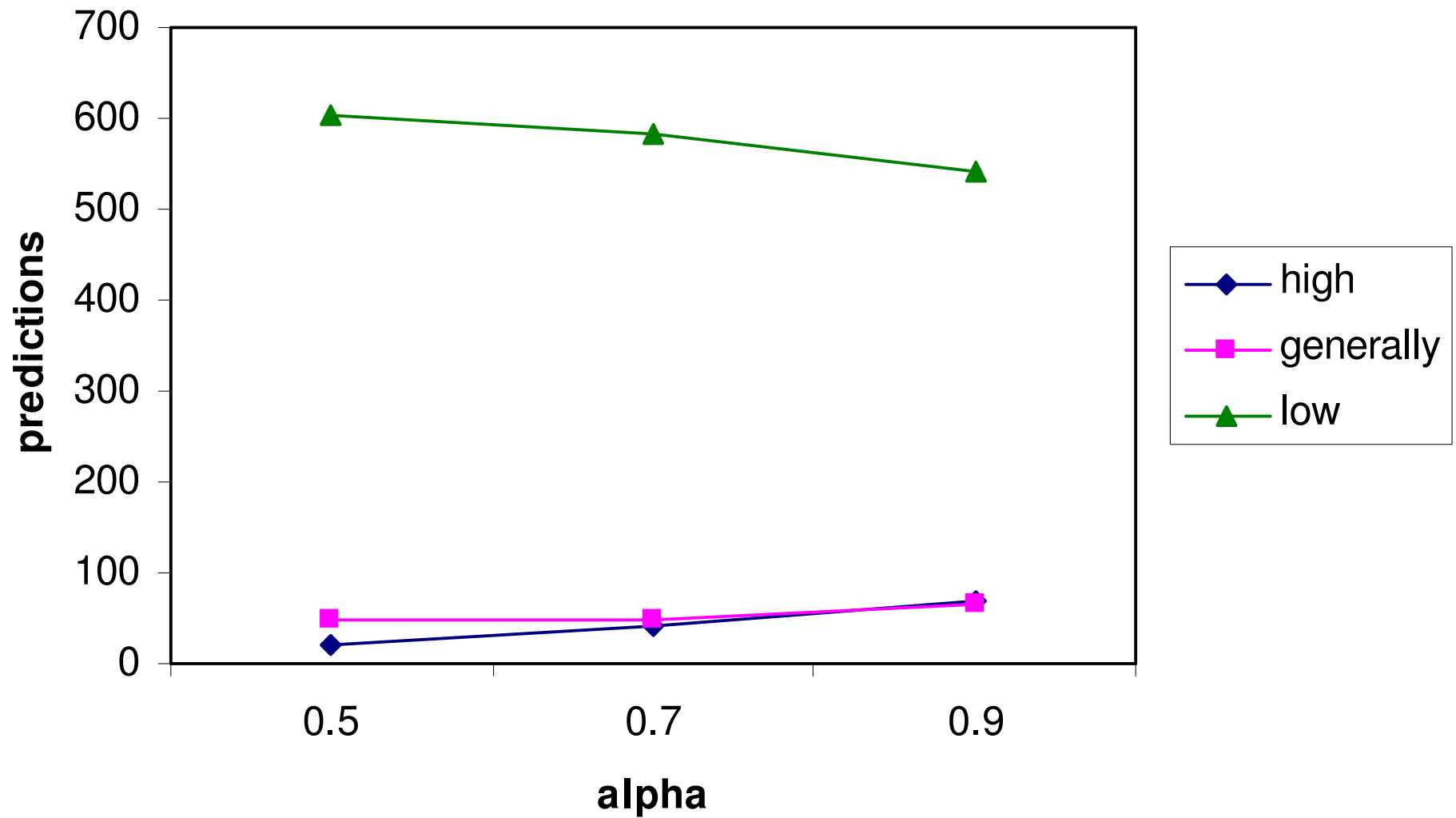




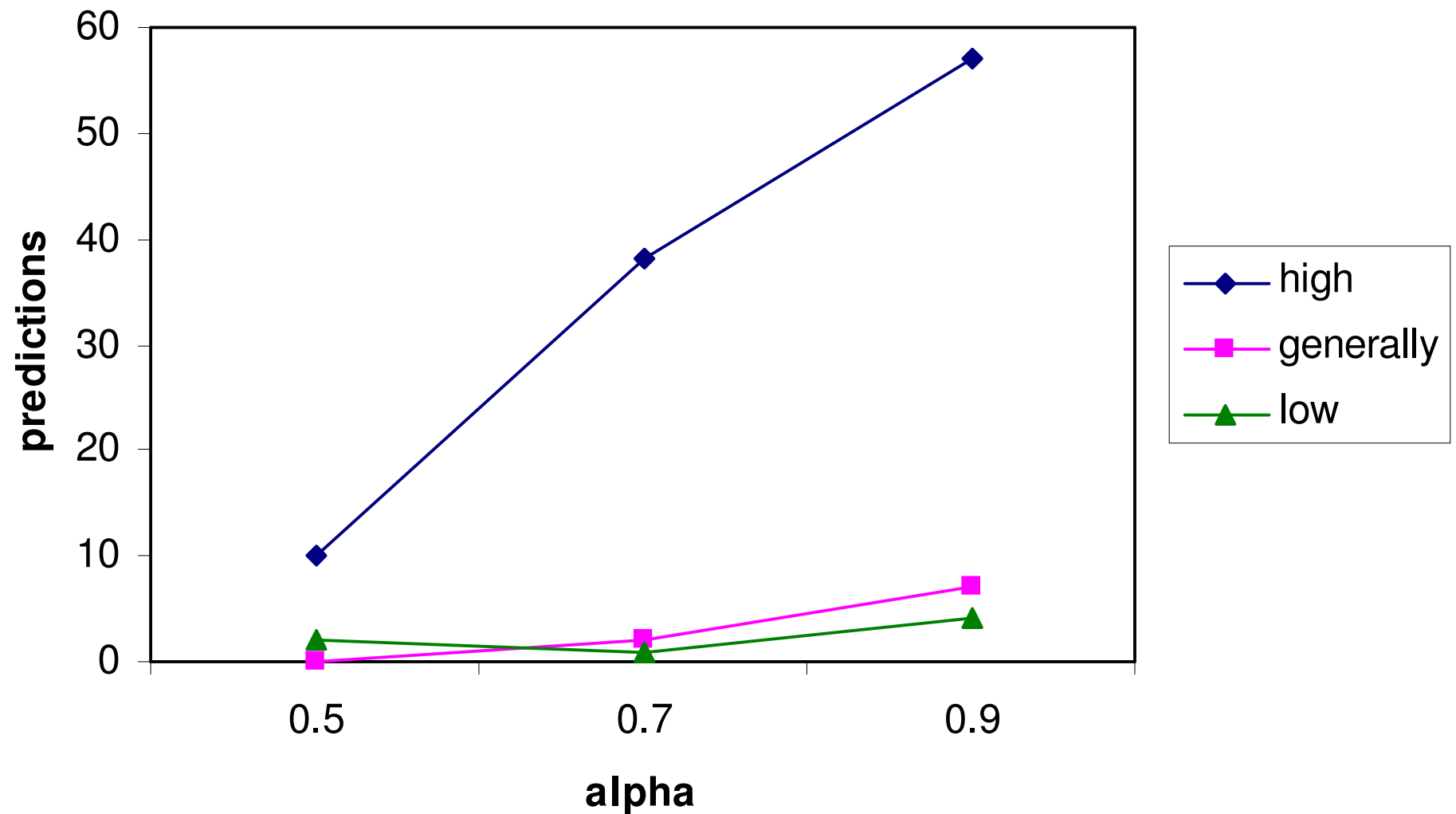
Discussion of task 2.1

- v Better GO identification with a smaller α
 - v Large number of terms not explicitly mentioned in the text
 - v Some correct sentences had less than 70% of the GO term's name
- v Better protein identification with a smaller α
 - v More sentences to filter by the protein's name presence
 - v About 50% of the predictions were incorrect because of the protein evaluation

Task 2.2 GO Evaluation



Task 2.2 Protein Evaluation





Discussion of task 2.2

- ✓ Better GO identification with a larger α
 - ✓ Terms with a larger piece of its name in the text were more accurate
 - ✓ Many terms mentioned but out of context
- ✓ Protein identification did not affect the results



Outline

v Introduction

v Method

v Results

v Conclusions



Conclusions

- v FiGO a novel method for identifying GO terms in unstructured text
- v Involving the information content of their names
- v FiGO is **fully automated**, i.e. it does not need human intervention
- v Despite the good score the results must be improved



Future Work

- v A more effective protein identification method would likely improve our results
- v The piece of text should be larger than a sentence
 - v Protein and term are normally in the same paragraph but not in the same sentence
 - v Number of occurrences in the document
- v Task 2.2 needs domain knowledge to filter terms out of context
 - v Using web resources
 - v (WeBTC presented at SAC2004)



<http://xldb.fc.ul.pt/rebil/>