university of groningen

GLAD: Groningen Lightweight Authorship Detection

PAN, Authorship verification, 2015

Manuela Hürlinmann, Benno Weck, Esther van den Berg, Simon Šuster, Malvina Nissim

The challenge

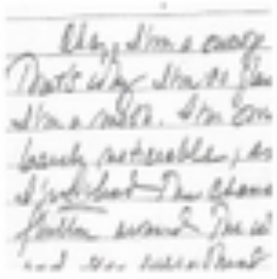
given: a **set** of **Known documents** written by the **same** Author A_K ,

given: **one Unknown document** written by an **unknown** Author A_U,

task: determine whether **A_U = A_K**

Gentle Tony

Fat Vinny

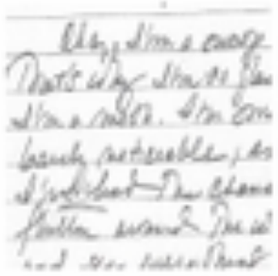


The Weasel

How can we recognise different authors?

Gentle Tony

Fat Vinny



The Weasel

How can we recognise different authors?

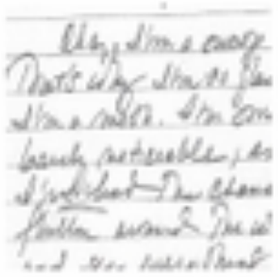
Unusual word choice?

Shorter sentences?

More complex grammar?

Gentle Tony

Fat Vinny



The Weasel

How can we recognise different authors?

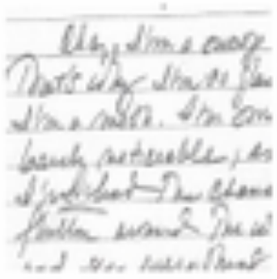
individual_vector(feat1, feat2…)

individual_vector(feat1, feat2…)

individual_vector(feat1, feat2…)

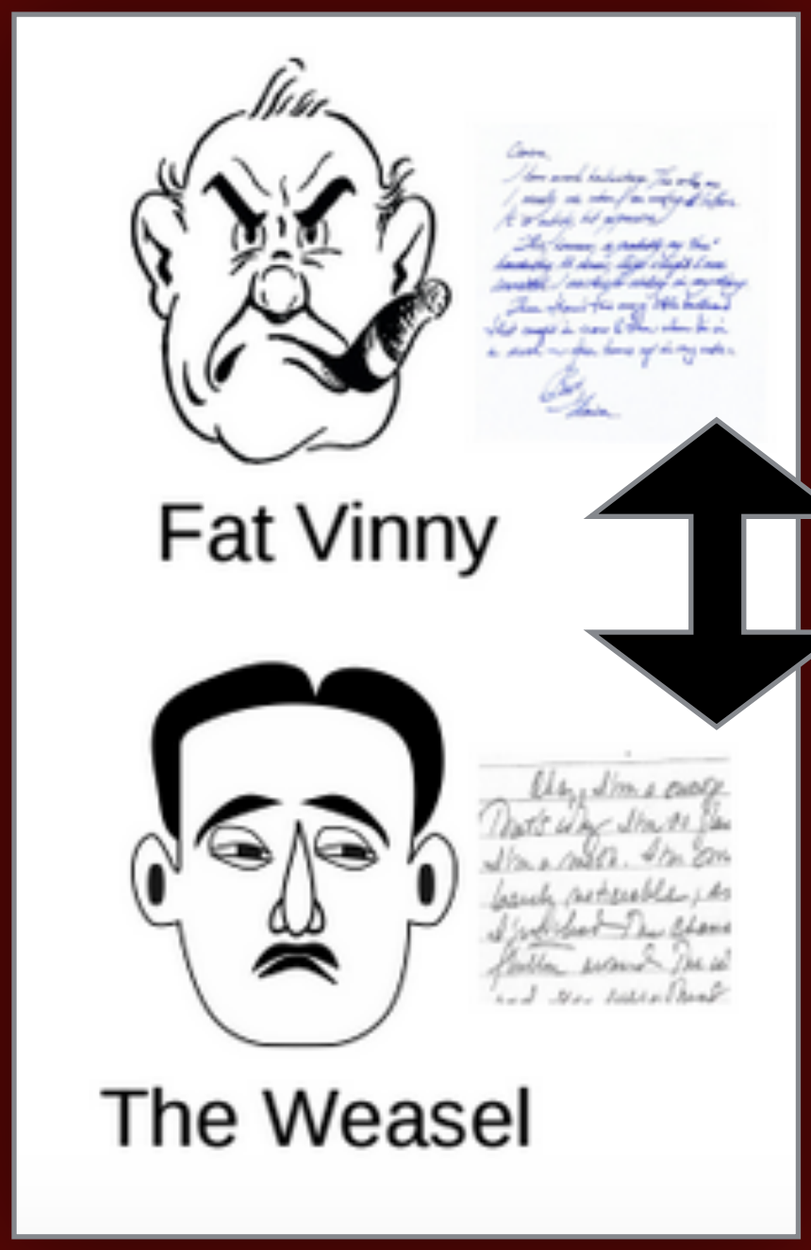
Gentle Tony

Fat Vinny



The Weasel

How can we then differentiate between authors?

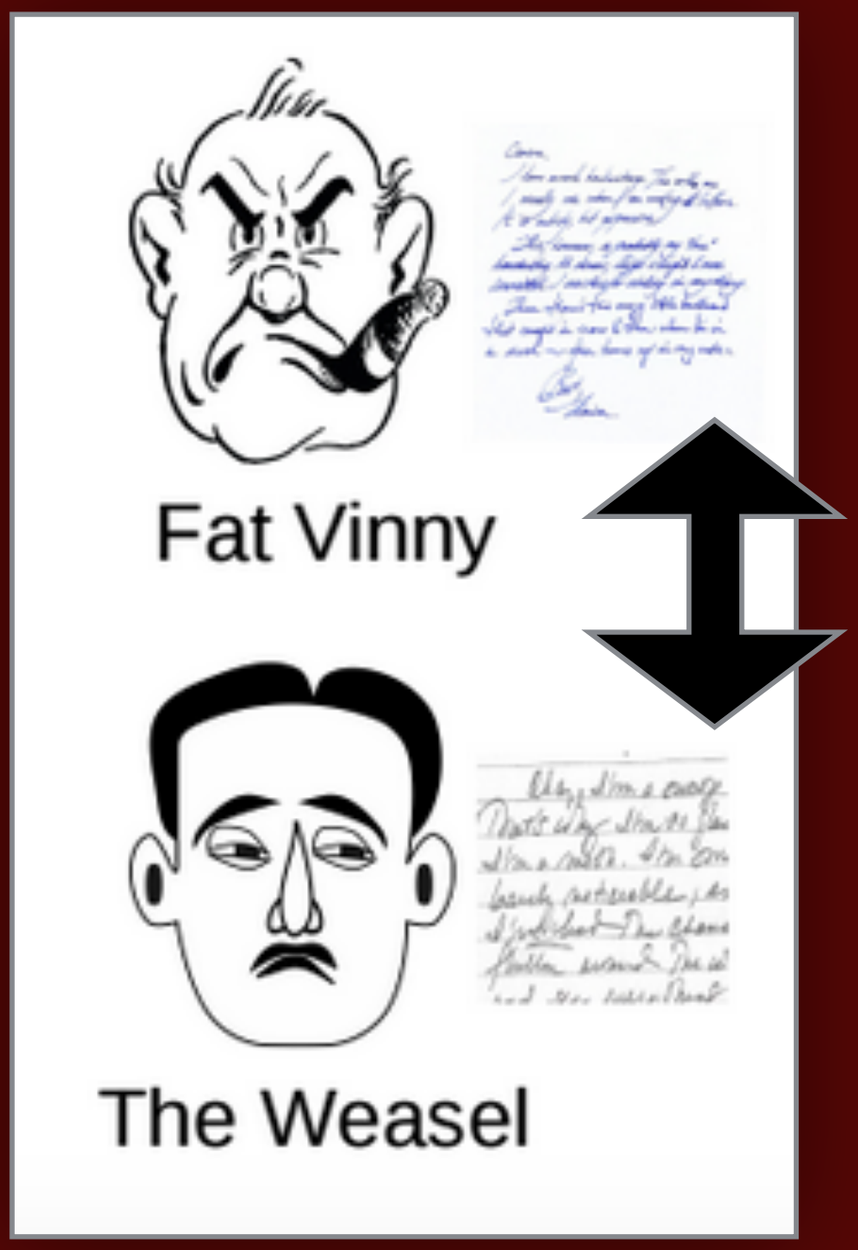


How can we then differentiate between authors?

Different word choice?

Different sentence length?

Different grammar?



How can we then differentiate between authors?

similarity_vector(feat1, feat2, …)

Our approach

- machine learning approach training on PAN (2015) data
- using SVM to do two-class classification task
- a set of features
- feature ablation studies to tune the system to each different language

The core aim

- A lightweight system!

The aim

training instance

training instance

training instance

Input in any language

The aim

training instance

training instance

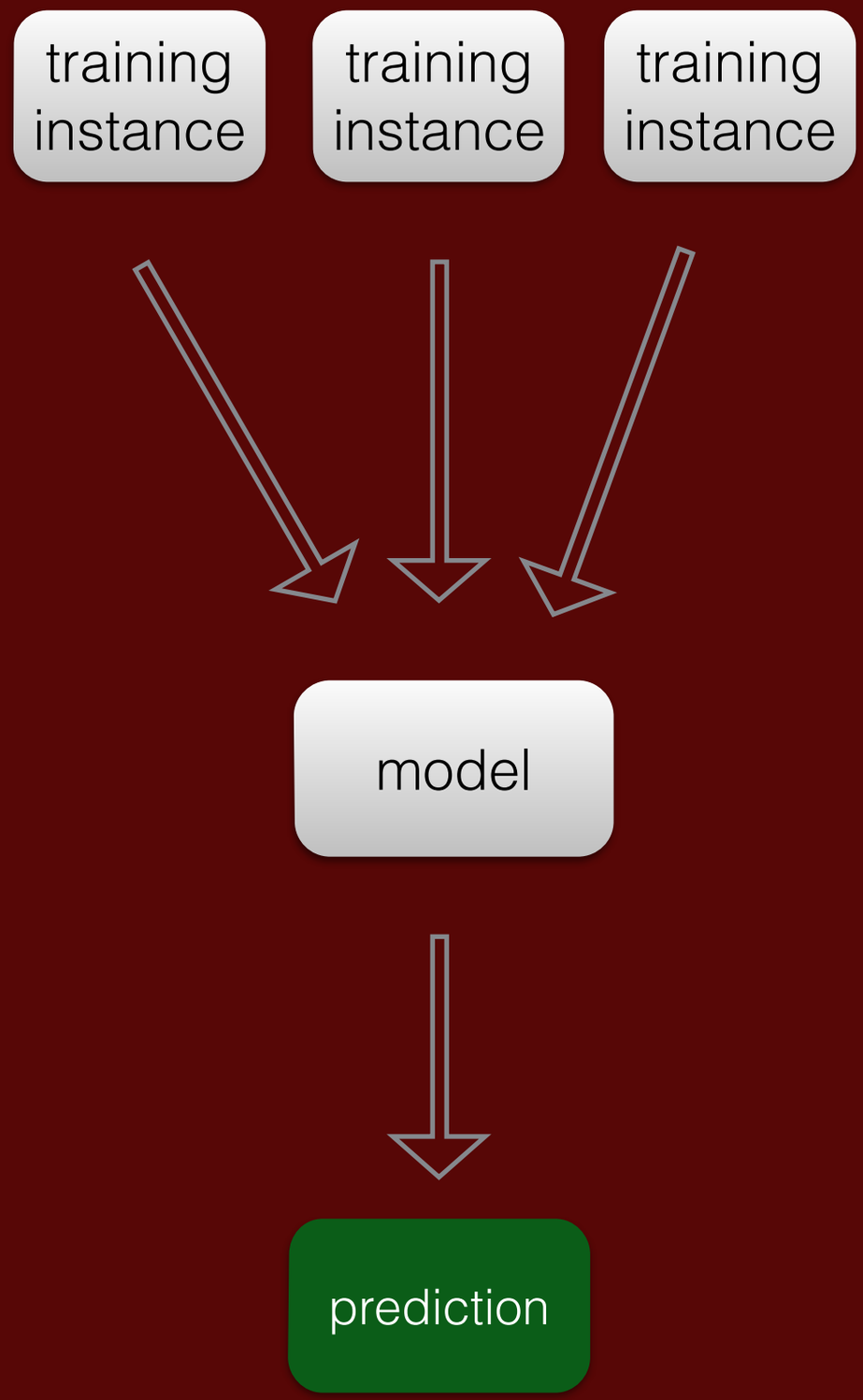
training instance

model

Input in any language

Features should be easy to extract

The aim

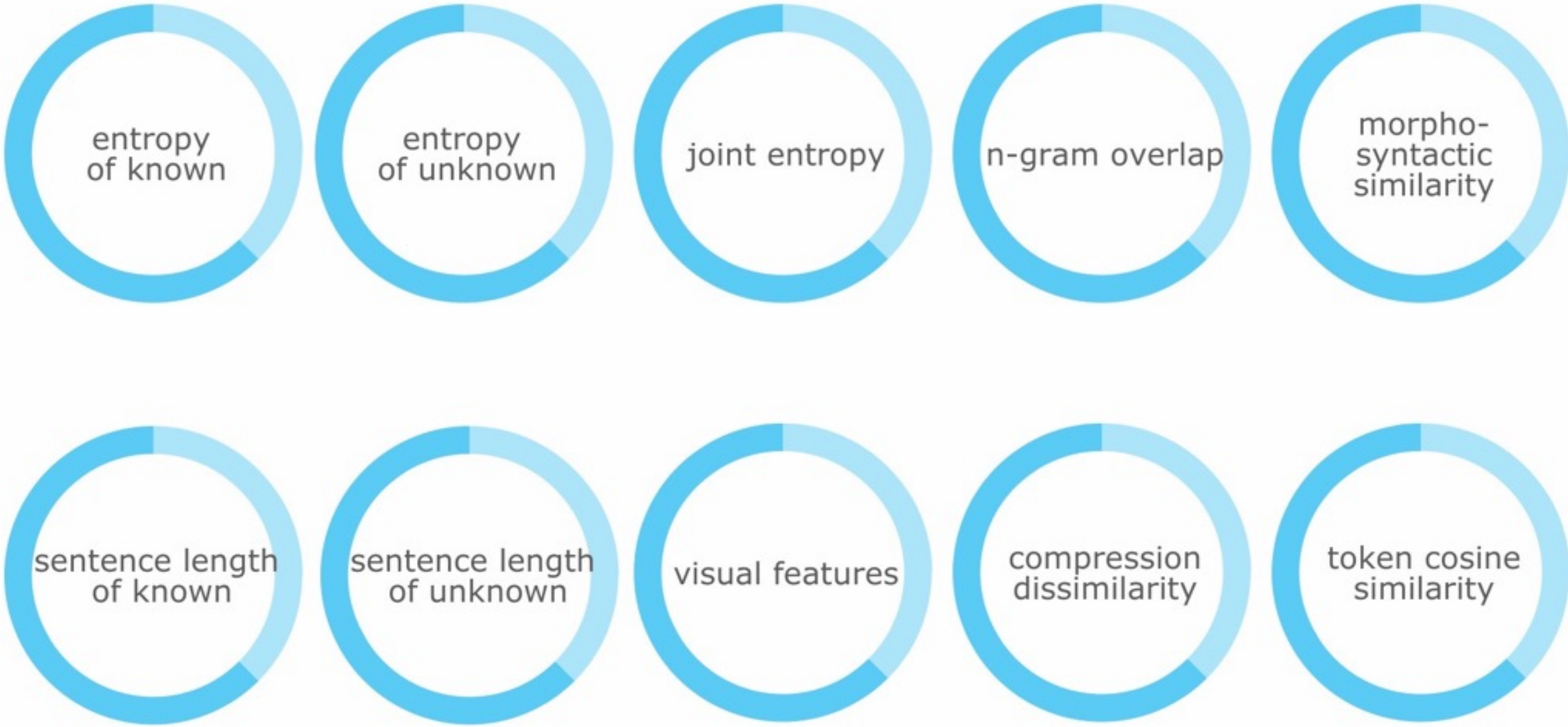


Input in any language

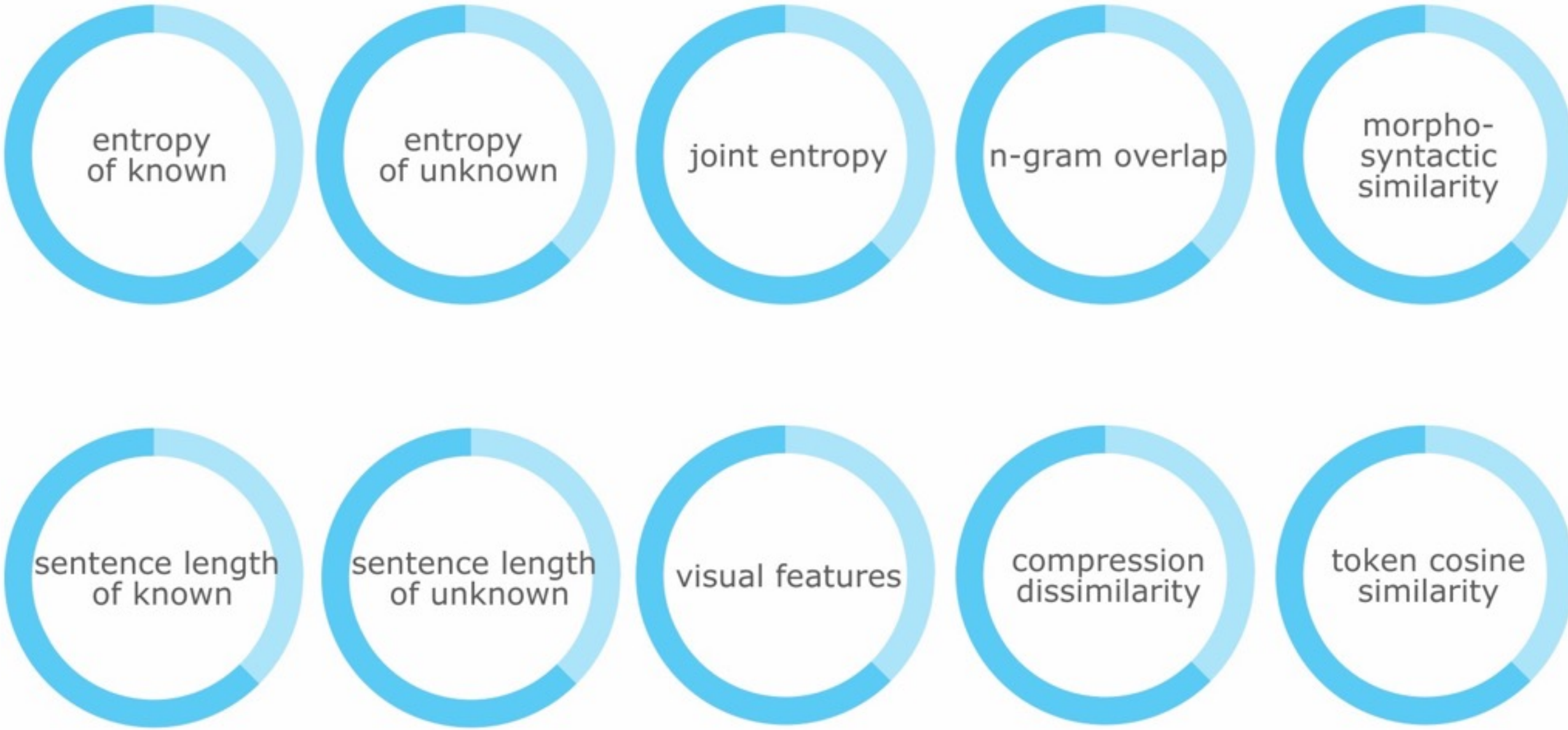
Features should be easy to extract

Training & Testing time should be fast

Our features



Our features



similarity_vector(entropy_of_known, visual_features, ...)

Our features

entropy of known

entropy of unknown

sentence length of known

sentence length of unknown

joint entropy

n-gram overlap

morpho-syntactic similarity

visual features

compression dissimilarity

token cosine similarity

To determine relevance: grouping

Our features

- entropy of known
- entropy of unknown
- sentence length of known
- sentence length of unknown
- joint entropy
- n-gram overlap
- morpho-syntactic similarity
- visual features
- compression dissimilarity
- token cosine similarity

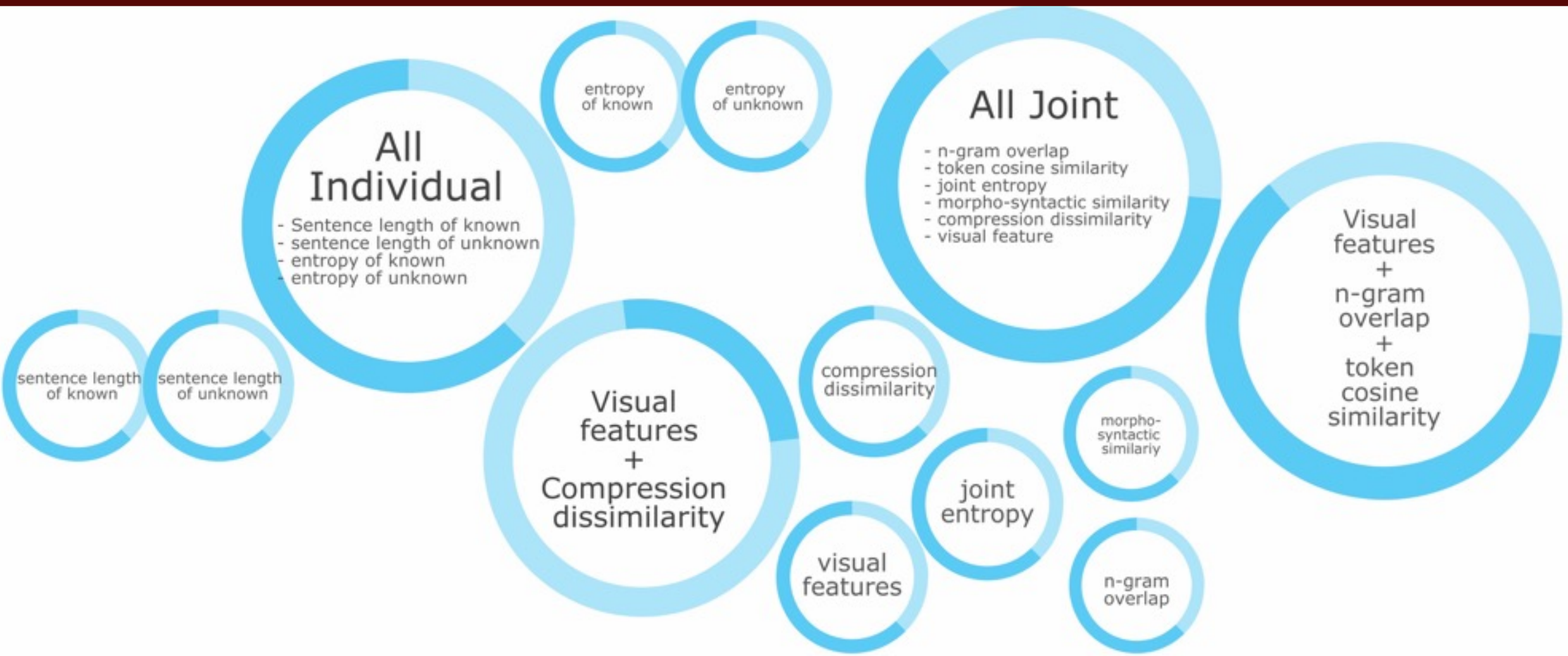
Individual → Vector_K(feat1,feat2)

Individual → Vector_U(feat1,feat2)

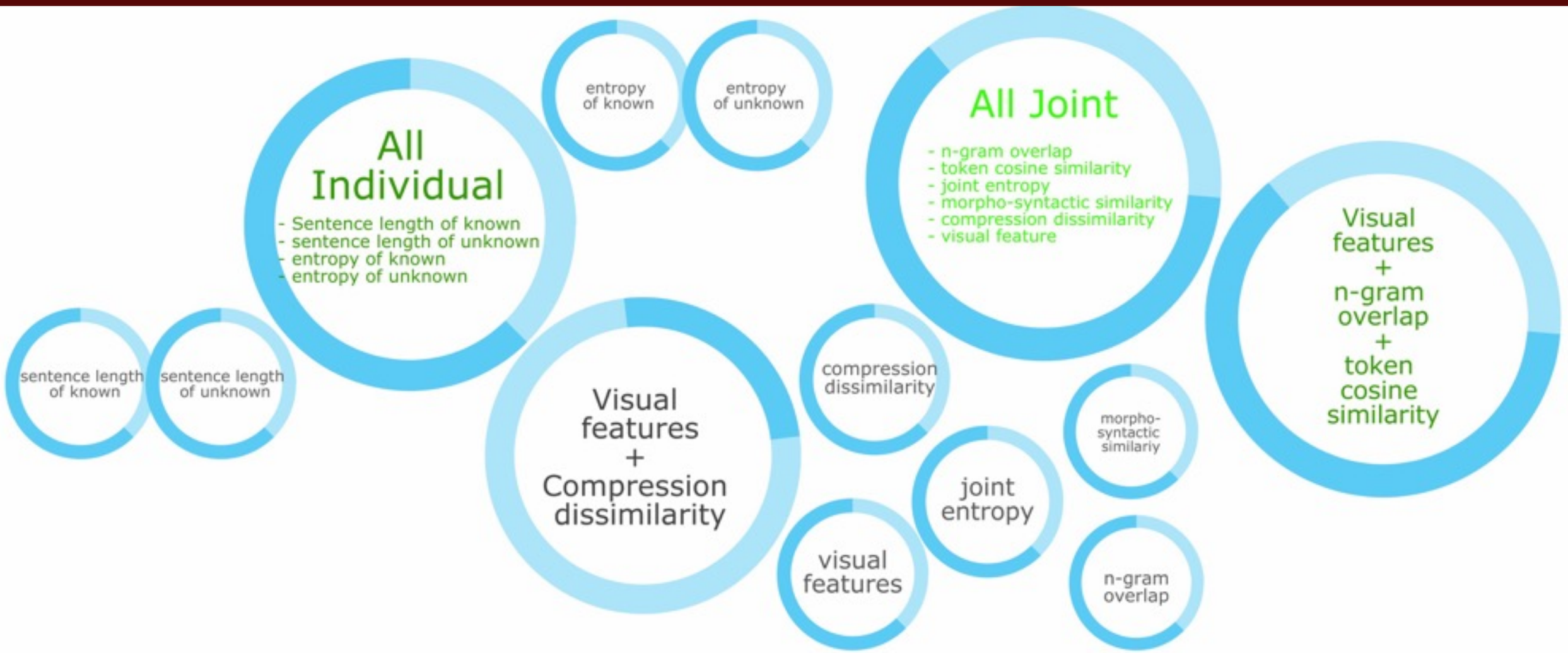
Joint → Vector_Joint(feat1,feat2)

Vector_K(feat1,feat2) - Vector_U(feat1,feat2) = Vector_Joint(feat1,feat2)

Comparing features

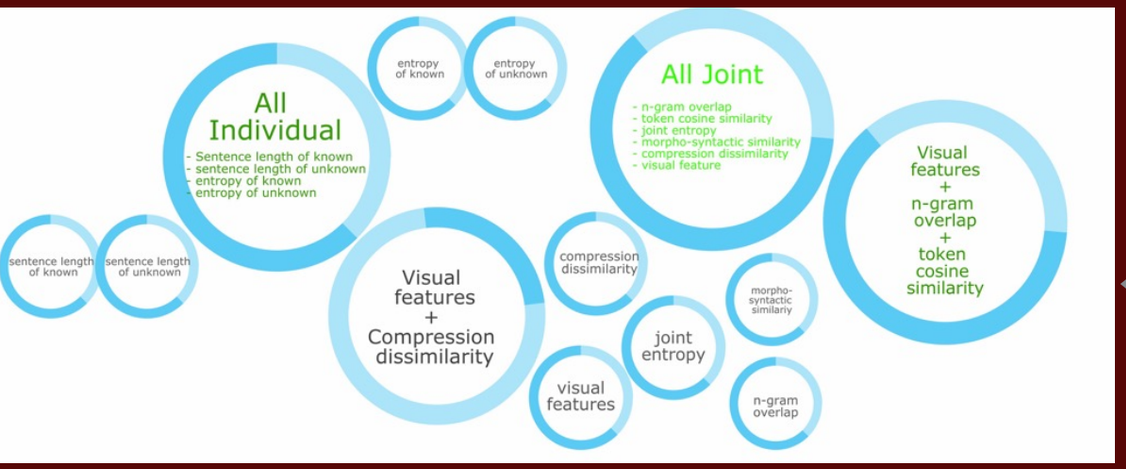


Comparing features



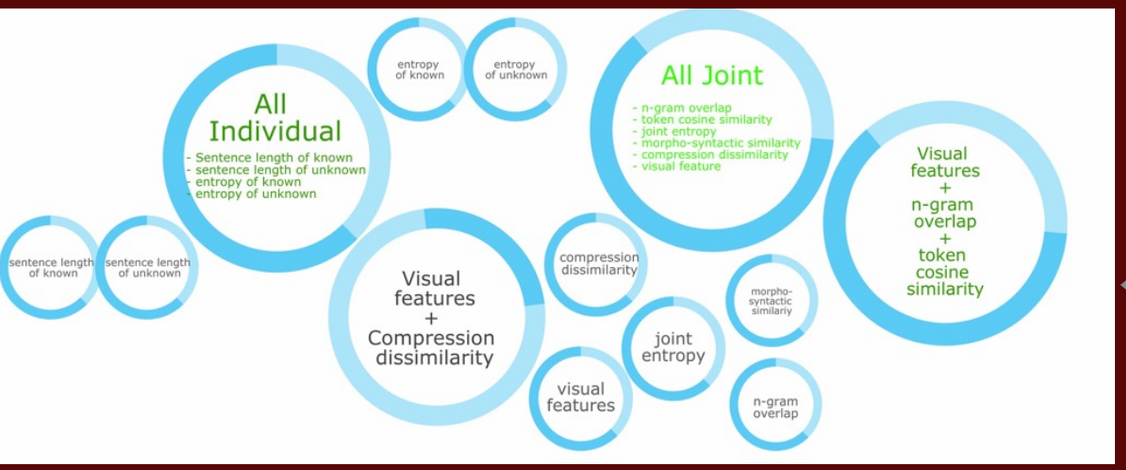
Results of ablation & single-feature experiments:

Helpful features



Side note:
Visual features

- Punctuation
- Line ending
- Letter case
- Ling length
- Block size

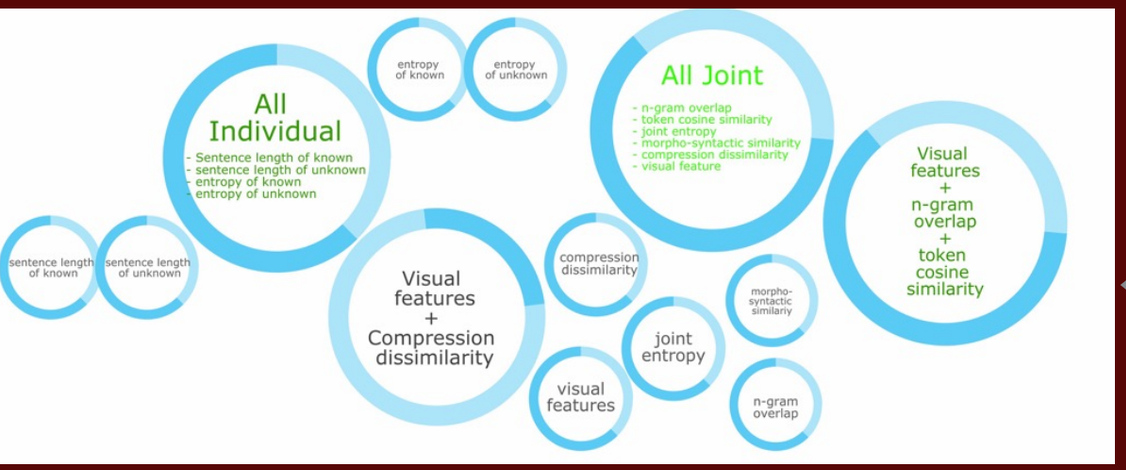


Side note:
Visual features

- Punctuation
- Line ending
- Letter case
- Ling length
- Block size

Con

- **Not a characteristic of the author**
- **Not a linguistic feature**



“Pa-pa, pa-pa, pa-pa!

Here, stop her. She’ll fall down.
Here, turn around. Walk this way.

Ma-ma, ma-ma, ma-ma;

Oh, I think you are a darling.

Mer-ry Christ-mas! Mer-ry Christmas.”

Side note:
Visual features

- Punctuation
- Line ending
- Letter case
- Ling length
- Block size

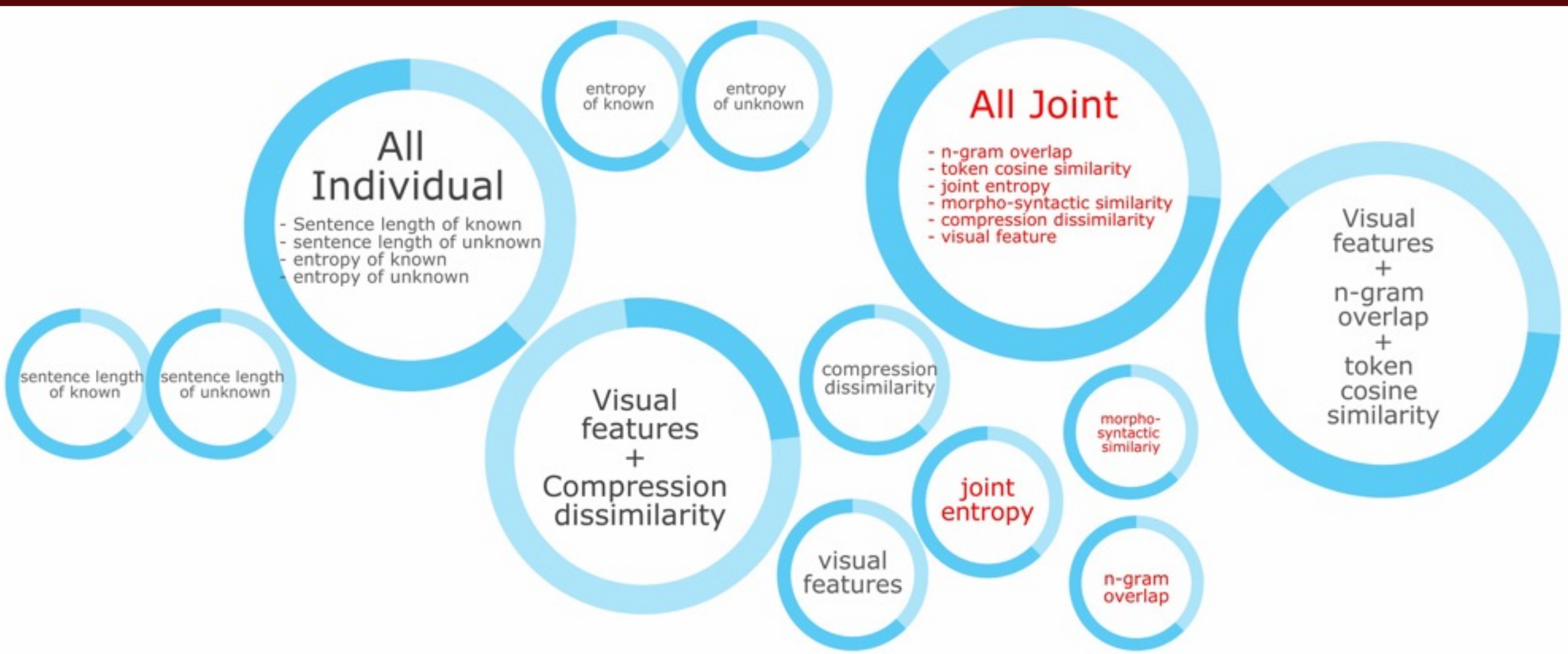
Con

- **Not a characteristic of the author**
- **Not a linguistic feature**

Pro

- **Can be author-specific for some genres**
- **If it works…**

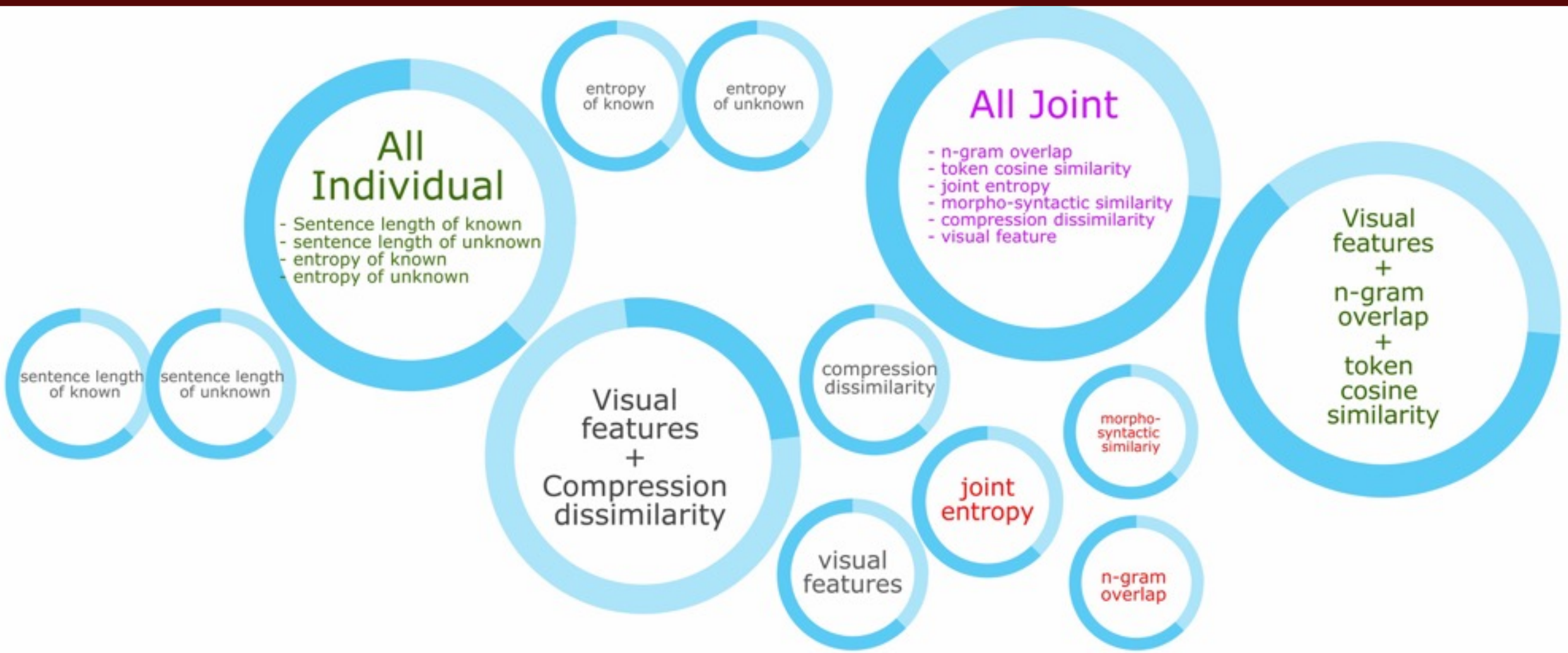
Comparing features



Results of ablation & single-feature experiments:

Harmful features

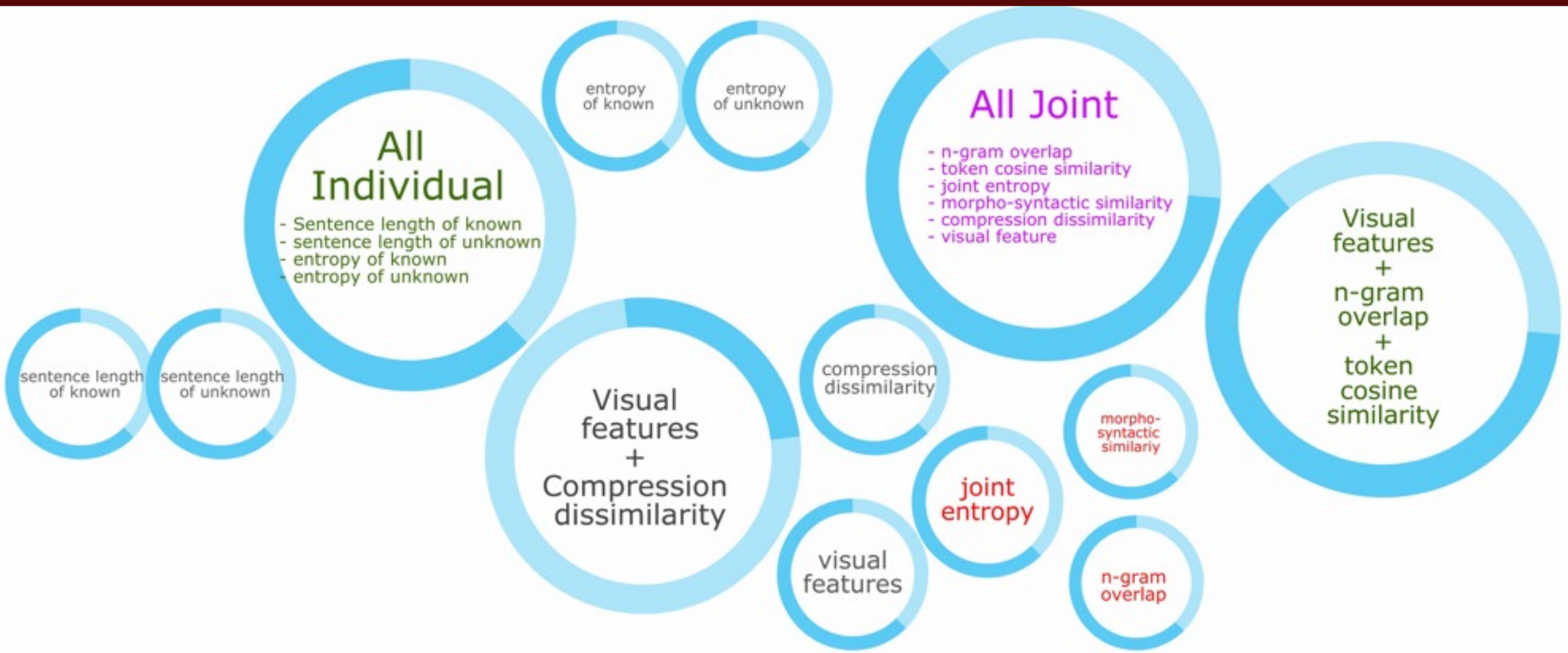
Comparing features



Results of ablation & single-feature experiments:

Features that are harmful, helpful, or helpful-depending-on-the-language

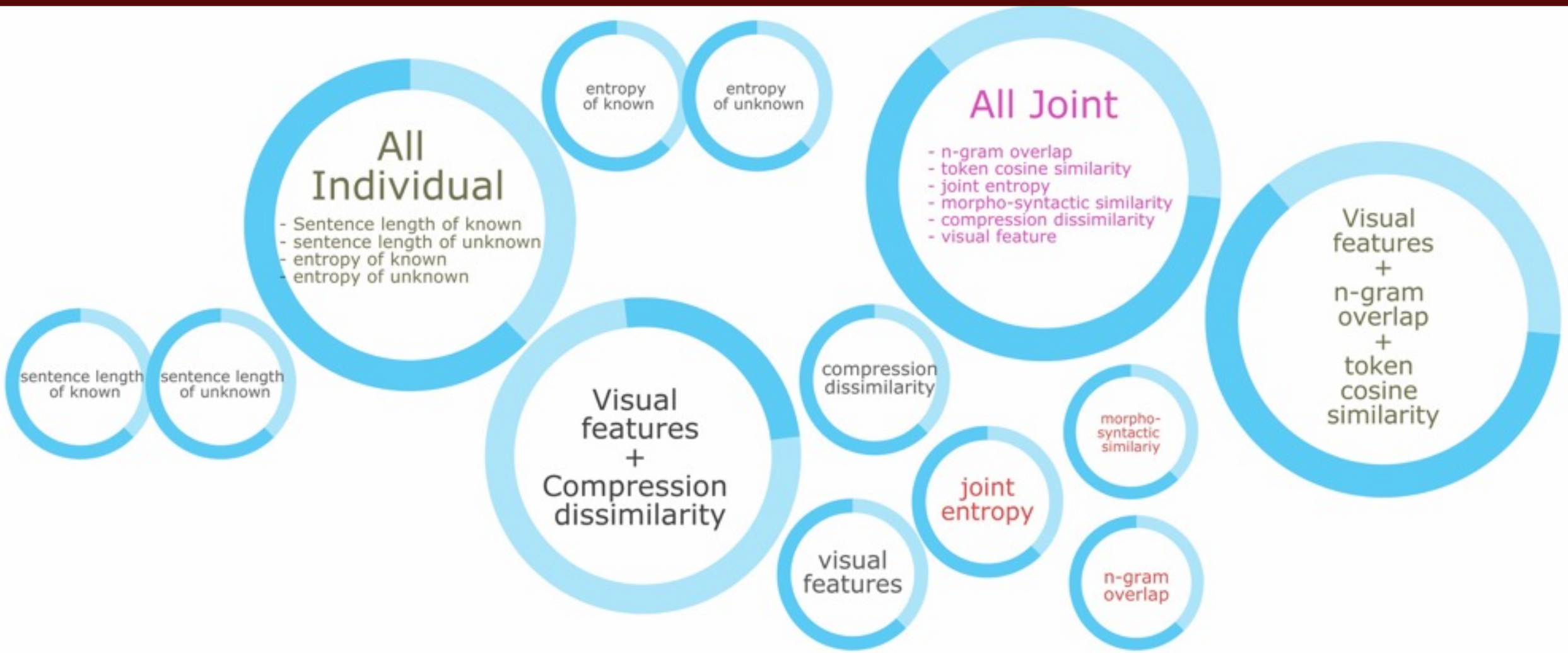
Comparing features



Results of ablation & single-feature experiments:

Features that are harmful, helpful, or helpful-depending-on-the-language

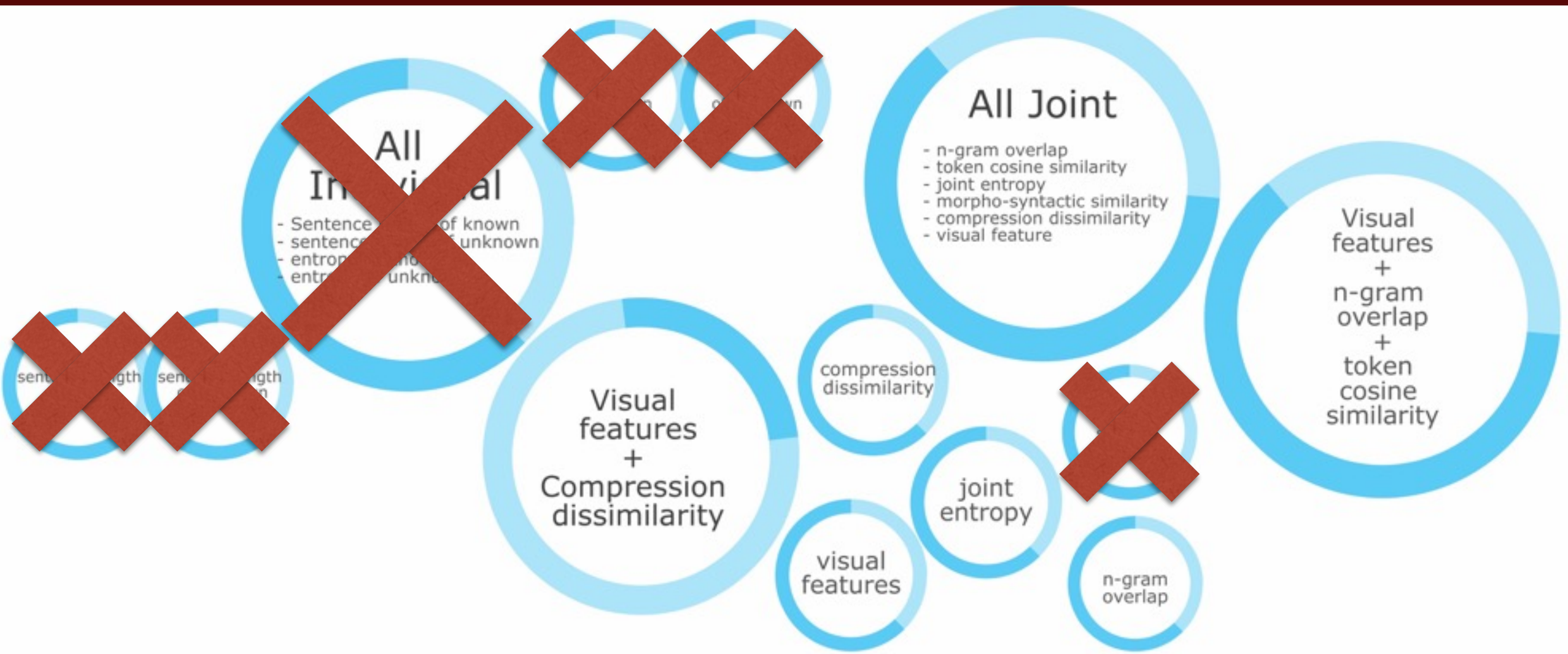
Comparing features



Results of ablation & single-feature experiments:

Differences are subtle

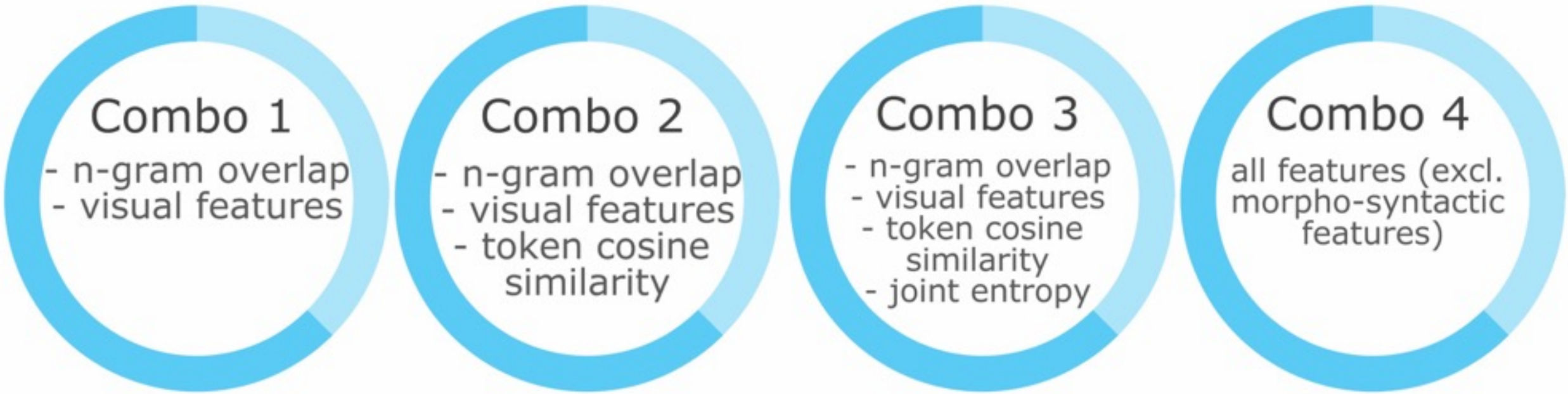
Comparing features



Results of ablation & single-feature experiments:

Differences are subtle

Resulting groups



Results

Language	Training	Test	Ranking
Dutch (`full set`)	.55	.62	3
English (`full set`)	.56	.41	7
Greek (`combo2`)	.54	.60	4
Spanish (`full set`)	.90	.54	5

Results

Language	Training	Test	Ranking
Dutch (`full set`)	.55	.62	3
English (`full set`)	.56	.41	7
Greek (`combo2`)	.54	.60	4
Spanish (`full set`)	.90	.54	5

- Simple similarity features work

Results

Language	Training	Test	Ranking
Dutch (`full set`)	.55	.62	3
English (`full set`)	.56	.41	7
Greek (`combo2`)	.54	.60	4
Spanish (`full set`)	.90	.54	5

- Simple similarity features work **in unison**

Results

Language	Training	Test	Ranking
Dutch (`full set`)	.55	.62	3
English (`full set`)	.56	.41	7
Greek (`combo2`)	.54	.60	4
Spanish (`full set`)	.90	.54	5

- Simple similarity features work **in unison** *independent of language (except greek)*

Results

Language	Training	Test	Ranking
Dutch (`full set`)	.55	.62	3
English (`full set`)	.56	.41	7
Greek (`combo2`)	.54	.60	4
Spanish (`full set`)	.90	.54	5

- Simple similarity features work **in unison** *independent of language (except greek)*
- System works fast (runtime av. 1 minute)

Final conclusion

GLAD

… is a light and fast language-independent system

… allows language adaptation done via feature selection

… involves innovative visual features which appear useful (especially for English data) and could be investigated further