Manual Alignment of Texts and Summaries in a Multi-document Corpus of News Articles

(Alinhamento Manual de Textos e Sumários em um *Corpus* Jornalístico Multidocumento)

[1,2]Verônica Agostini

[1,3]Renata Tironi de Camargo

[1,3]Ariani Di Felippo

[1,2]Thiago A. S. Pardo

1 Núcleo Interinstitucional de Linguística Computacional (NILC)
2 Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP)
3 Departamento de Letras (DL), Universidade Federal de São Carlos (UFSCar)

Schedule

- Context
- Alignment Background
- The CSTNews Corpus
- The CTSNews Alignment
  - General Rules
  - Specific Rules
- Results
- Final Remarks

Context

- **Multi-document Summarization (MDS)**

  "Multi-document summarization is the automatic production of a unique summary from a collection of texts on a same topic." (Mani, 2001)

  - A lot of information available

- **Alignment**
  - Among text segments with overlapping content

- **Goals**
  - Linguistic analysis of MDS
  - Rules and models for MDS

Example

Document 1

[1] As piores enchentes dos últimos 60 anos no Reino Unido deixaram milhares de britânicos desabrigados, sem abastecimento de água ou sem energia elétrica. [2] O país sofre com as chuvas que caem desde junho. [3] Na sexta-feira, choveu 12 centímetros em algumas regiões, e há previsão de mais tempestades hoje.

[4] O ministro do Ambiente, Hilary Benn, afirmou que a emergência está "longe de acabar, e mais inundações são prováveis". [5] Benn disse que oito alertas de enchentes severas já foram dados. [6] "Cerca de 10 mil casas já estão ou podem ser inundadas."

[7] O rio Severn, o maior do país, está cinco metros acima do nível normal de verão. [8] O degelo da neve também influi no aumento do nível dos rios. [9] "Nunca vimos enchentes dessa altura", afirmou um porta-voz da Agência Ambiental. [10] O nível máximo de inundação no Reino Unido foi registrado em 1947.

[11] Cerca de 350 mil britânicos do sudoeste do país podem ficar sem abastecimento de água devido à ameaça de mais transbordamento dos rios Severn e Tâmisa.

Example

Document 2

[1] A chuva torrencial que atinge o Reino Unido encobriu estradas e milhares de pessoas estão sem fornecimento de eletricidade e de água potável em decorrência da pior enchente nos últimos 60 anos no país, segundo informou nesta segunda-feira, 23, a rede de televisão BBC.

[2] Os dois maiores rios do Reino Unido, Severn e Tâmisa, ameaçam transbordar nesta segunda, agravando ainda mais a situação nas regiões centro e sul da Inglaterra, que vêm sendo castigadas por inundações desde a última sexta-feira.

[3] Serviços de meteorologia (...).

[4] O nível da água do rio Severn (...).

[5] Já o rio Tâmisa, que está com seu leito no limite, pode transbordar durante a próxima madrugada caso as chuvas continuem, informou a Agência de Meio-Ambiente do Reino Unido.

[6] O primeiro-ministro Gordon Brown está em Gloucestershire, o condado mais afetado pelas enchentes, que deixaram 150 mil casas sem água.

[7] Cerca de 43 mil residências também estão sem eletricidade, depois que uma estação na cidade de Gloucester teve de ser desligada quando o nível da água começou a subir. [8] Em Oxford, 1.500 pessoas foram evacuadas para o maior estádio de futebol da cidade.

[9] O condado de Worcestershire também foi castigado pelas enchentes, onde o nível da água atingiu 1,82 metro e cerca de quatro mil pessoas estão sem água.

[10] Desde sexta-feira, centenas de pessoas na Inglaterra foram resgatadas de suas casas, no que a Força Aérea britânica chamou de "a maior operação de resgate em tempos de paz". **[11]** Nas piores áreas, o Exército pediu ajuda à Aeronáutica, que enviou helicópteros para resgatar mais de cem moradores de suas casas.

[12] No fim de semana (...).

[13] Fontes da Agência de Meio Ambiente (...).

[14] A Associação de Seguradores Britânicos estima que a conta para cobrir os estragos provocados pelas enchentes dos meses de junho e julho no país chegue a 2 bilhões de libras (R$ 7,6 bilhões).

Example

Multi-document Summary of 2 documents

[1] A chuva torrencial que atinge o Reino Unido encobriu estradas e milhares de pessoas estão sem fornecimento de eletricidade e de água potável em decorrência da pior enchente nos últimos 60 anos no país.

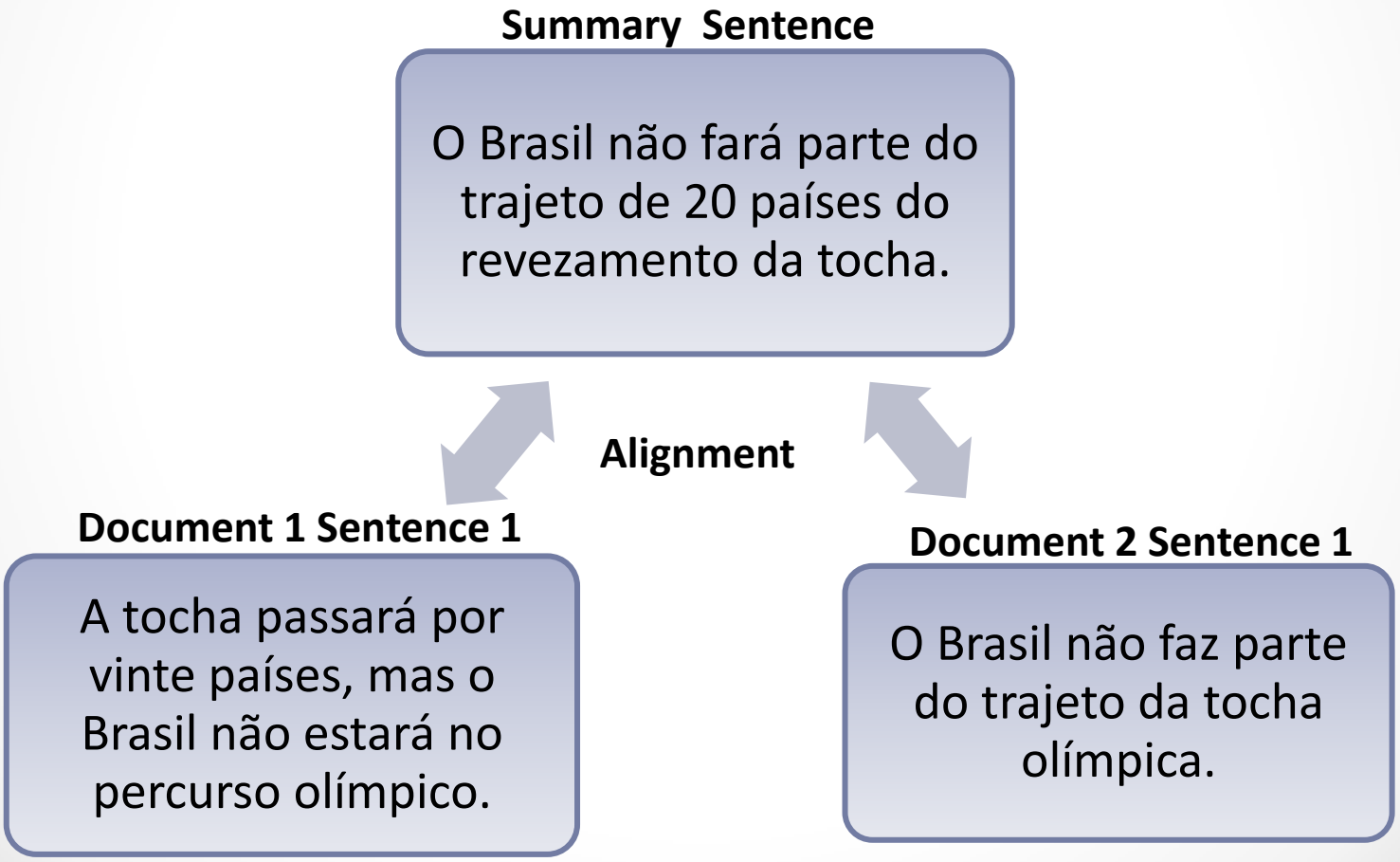
[2] Na sexta-feira, choveu muito acima do esperado e há previsão de mais tempestades hoje. **[3]** Os dois maiores rios do Reino Unido, Severn e Tâmisa, ameaçam transbordar nesta segunda, agravando ainda mais a situação. **[4]** O degelo da neve também influi no aumento do nível dos rios.

[5] Desde sexta-feira, centenas de pessoas na Inglaterra foram resgatadas de suas casas, no que a Força Aérea britânica chamou de "a maior operação de resgate em tempos de paz".

[6] Estima-se que serão necessários cerca de 2 bilhões de libras para cobrir os estragos.

Example

- Alignment type: 1 - 2

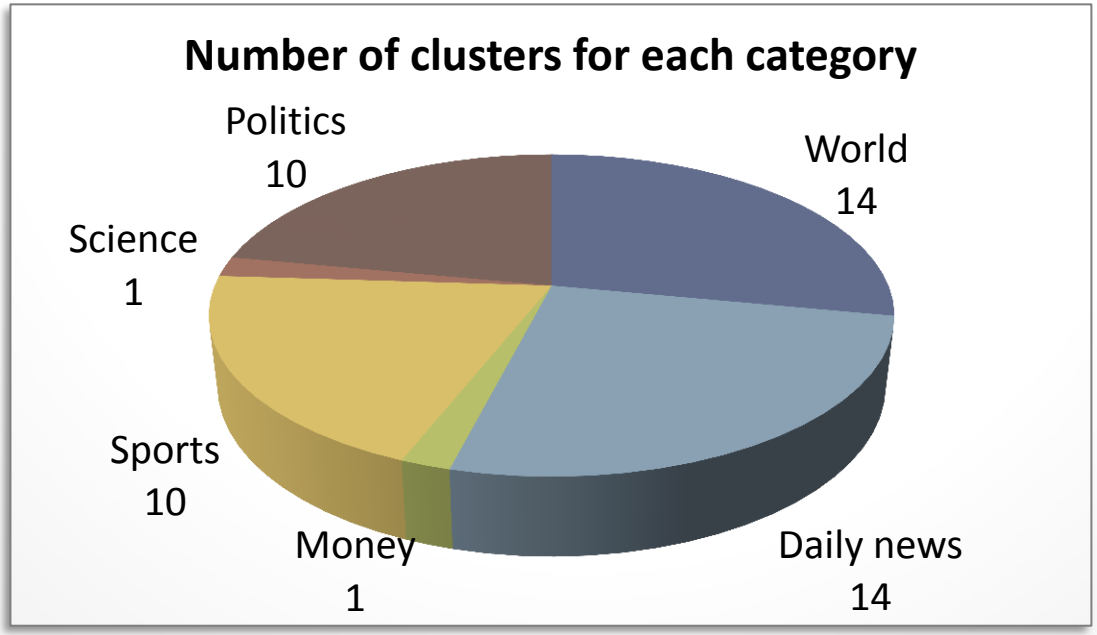


The CSTNews Corpus

- 50 clusters of news texts
  - Each cluster has:
    - from 2 to 3 texts in Brazilian Portuguese (140 texts)
    - automatic and manual multi-document summaries
    - 42 sentences on average (10 to 89) – documents
    - 7 sentences on average (3 to 14) – manual summaries
  - And also:
    - Discourse annotation (RST, CST, aspects)
    - Temporal annotation (HAREM guidelines)
    - WSD for nouns
    - Subtopic/topic segmentation
    - Etc.

The CSTNews Corpus

- The documents are from 5 online news agencies:
  - ***Folha de São Paulo, Estadão, Jornal do Brasil, O Globo*** and ***Gazeta do Povo***
- The clusters are distributed according to 6 categories:
  - **politics, world, daily news, money, sports** and **science**



The CSTNews Alignment

- 2 annotators
- Initial criteria
  - Sentential level
  - Content overlap

Summary Sentence	Document Sentence
Vários moradores e turistas nas regiões, inclusive brasileiros, foram retirados dos locais, enquanto outros estão **se preparando para a passagem do furacão**.	Na Jamaica, muitos **estocaram alimentos, água, lanternas e velas.**

The CSTNews Alignment

- Training
  - Selection of 2 clusters randomly
  - Decisions of agreement

- Annotation
  - From 1h to 2h a day
  - 2 months on average
  - Creation of 8 rules

- Agreement
  - Selection of 5 clusters randomly, considering its section

General Rules

- **Rule 1**
  - Align based on content overlap

Summary Sentence

17 pessoas morreram após a queda de um avião na República Democrática do Congo.

Document Sentence

Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.

General Rules

- **Rule 2**
  - Align based on main information overlap

Summary Sentence

O presidente também **afirmou** que o critério para os municípios e Estados contemplados com obras é técnico.

Document Sentence

Lula disse que o critério para o investimento nas cidades será técnico, não partidário.

General Rules

- **Rule 2**
  - Align based on main information overlap

Summary Sentence

Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, **descobriram** um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que giram um ao redor do outro, denominado Oph 162225-240515, o primeiro planemo duplo.

Document Sentence

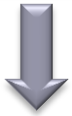
Os pesquisadores Ray Jayawardhana e Valentin D. Ivanov **informam** a descoberta na edição de quinta-feira do serviço online Science Express, mantido pela revista Science.

General Rules

- **Rule 3**
  - Align based on secondary information overlap

Summary Sentence

Renan é alvo de um processo por quebra de decoro acusado de receber recursos da construtora Mendes Junior para **pagamento de despesas pessoais, como aluguel e pensão para a jornalista Mônica Veloso**, com quem tem uma filha.



Document Sentence

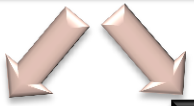
Isso permitiria que os peritos da Polícia Federal pudessem trabalhar durante o período de descanso dos senadores e, no retorno das férias, apresentarem um relatório detalhado sobre o conjunto de documentos - notas fiscais, recibos de vacinação, extratos bancários, guias de transporte de animais - que o senador apresentou para justificar o **pagamento da pensão informal à jornalista Mônica Veloso**.

General Rules

- **Rule 4**
  - Align all the overlapping content

Summary Sentence

Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que **giram um ao redor do outro**, denominado Oph 162225-240515, o primeiro planemo duplo.



Document 1 Sentence 1

Astrônomos do Observatório Europeu Austral, localizado no Chile, anunciaram a descoberta de uma dupla de planetas errantes (sem estrela-mãe) que **giram ao redor deles mesmos** e que vagam livremente pelo espaço.

Document 2 Sentence 1

O fato extraordinário é que **ele não gira em volta de uma estrela, mas em torno de outro corpo frio** com o dobro de sua massa.

Specific Rules

- **Rule 5**
  - Align even when there is contradictory numerical data

Summary Sentence

Às **9h**, a cidade tinha oito pontos de alagamento, sendo dois intransitáveis.

Document Sentence

O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às **9h30** desta segunda-feira.

Specific Rules

- **Rule 6**
  - Align even when there are different levels of generalization

Summary Sentence

A Companhia de Engenharia de Tráfego (CET) anunciou que o índice de congestionamento era de **54 quilômetros** às 8h, **113 km** às 9h e **110 km** meia hora depois, valores bem acima das médias para os horários, que eram de **36, 82 e 76 quilômetros** respectivamente, mas não havia registro de acidentes graves, apesar de haver feridos.

Document Sentence

Com o asfalto molhado, o trânsito ficou mais lento e **o congestionamento ficou o dobro da média.**

Specific Rules

- **Rule 7**
  - Align even when there are different levels of assertiveness

Summary sentence

As ações <u>são atribuídas</u> à facção criminosa Primeiro Comando da Capital (PCC), que já comandou outros ataques em duas ocasiões.

Document sentence

As ações criminosas <u>podem ter sido ordenadas</u> pelos líderes do Primeiro Comando da Capital (PCC), que haviam prometido retomar os ataques no Estado de São Paulo no Dia dos Pais, no próximo domingo.

Specific Rules

- **Rule 8**
  - **Don't** align when one express "whole" and the other a "part"

Summary Sentence

Somente neste ano, o senador **se internou por três vezes** no Incor.

Document Sentence

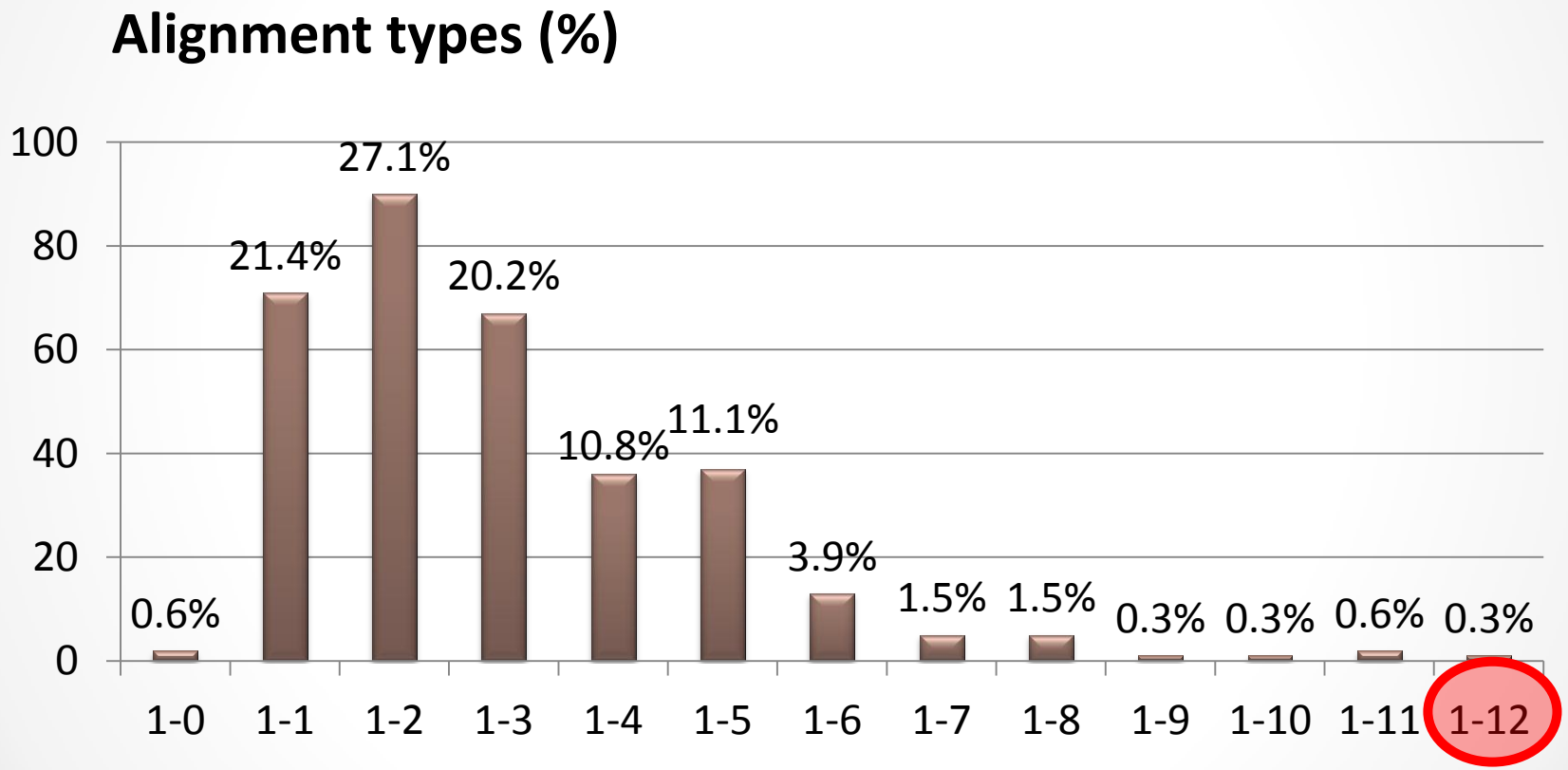
Em abril, o senador foi internado no InCor com insuficiência cardíaca.

Results

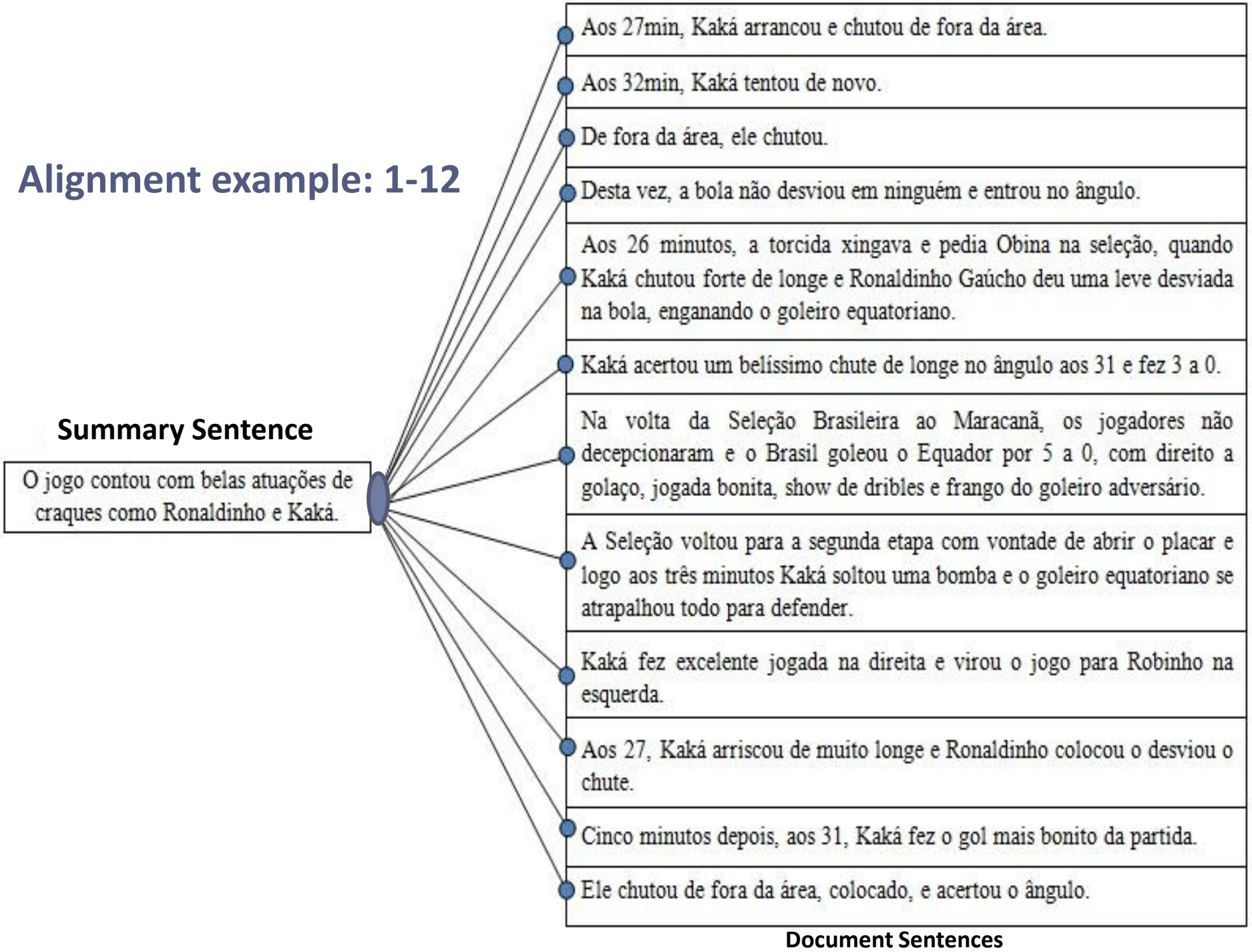
Amount of Alignments	Alignment types												
	1-0	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12
	2	**71**	**90**	**67**	36	37	13	5	5	1	1	2	1

Summary Sentences	Document Sentences	Aligned Sentences	Agreement (kappa)
331	2067	877	0.831

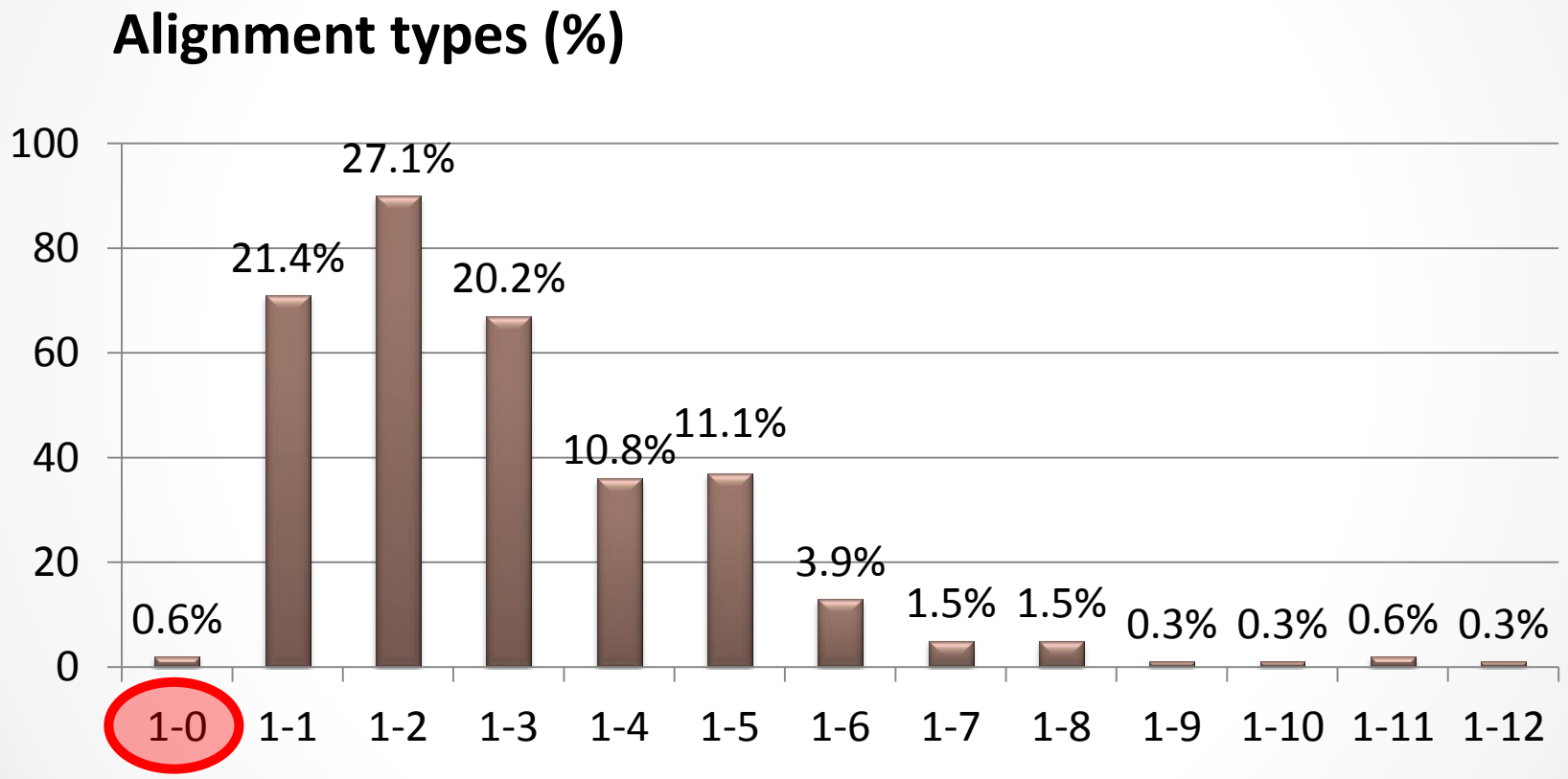
Results


Alignment types (%)
100
80
60
40
20
0
0.6%
21.4%
27.1%
20.2%
10.8%
11.1%
3.9%
1.5%
1.5%
0.3%
0.3%
0.6%
0.3%
1-0
1-1
1-2
1-3
1-4
1-5
1-6
1-7
1-8
1-9
1-10
1-11
1-12

Alignment example: 1-12

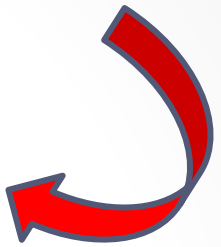


Results


Alignment types (%)
100
80
60
40
20
0
0.6%
21.4%
27.1%
20.2%
10.8%
11.1%
3.9%
1.5%
1.5%
0.3%
0.3%
0.6%
0.3%
1-0
1-1
1-2
1-3
1-4
1-5
1-6
1-7
1-8
1-9
1-10
1-11
1-12

Alignment example: 1-0

Inferred Sentence



Summary Sentence

Neste domingo, o esporte brasileiro alegrou a torcida verde-amarelo.

Document 1 Sentence 1

Foi um domingo especial e inesquecível para o esporte brasileiro.

Document 1 Sentence 2

O domingo verde-amarelo começou com a conquista do heptacampeonato da Liga.

Results

- **Representation of the results (XML)**

```
<align SENT="1">
    <DOC="D1_C31_Folha.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
</align>

<align SENT="2">
    <DOC="D1_C31_Folha.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="6" TYPE="none" JUDGE="veronica"/>
</align>

<align SENT="3">
    <DOC="D1_C31_Folha.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
</align>
```

Final Remarks

- Difficulties
  - Domain knowledge
    - **Sport texts**: specific rules
    - **Political texts**: specific terms

Example

Sport section: “pole vault” competition at the Pan American Games

Document Sentence

(...) a brasileira conseguiu o ouro na segunda tentativa.

Summary Sentence

(...) a brasileira conseguiu o ouro em três tentativas.

Main References

- Cardoso, P.C.F.; Maziero, E.G.; Castro Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting, pp. 88-105. October 26, Cuiabá/MT, Brazil.

- Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, v. 22, n. 2, pp. 249-254.

- Mani, I. (2001). *Automatic Summarization. Natural Language Processing Vol. 3*. Amsterdam/Philadelphia: John Benjamins Publishing Company. [Centro de Linguística da Universidade do Porto, Cota: N/35], 285 pp.

Manual alignment of texts and summaries in a multi-document corpus of news articles

(Alinhamento manual de textos e sumários em um *corpus* jornalístico multidocumento)

[1,2]Verônica Agostini

[1,3]Renata Tironi de Camargo

[1,3]Ariani Di Felippo

[1,2]Thiago A. S. Pardo

1 Núcleo Interinstitucional de Linguística Computacional (NILC)
2 Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP)
3 Departamento de Letras (DL), Universidade Federal de São Carlos (UFSCar)