Knowledge Extraction for Professionals from Financial Documents

Core AI, Goldman Sachs R&D

Disclaimer:

"Views and opinions expressed are for informational purposes only. They do not constitute a recommendation by Goldman Sachs to buy, sell or hold any security. No part of this material may, without Goldman Sachs's prior written consent, be (i) copied, photocopied or duplicated in any form, by any means, or (ii) distributed to any person that is an employee, officer, director, or authorized agent of the recipient.
Copyright Note: 2020 Goldman Sachs. All rights reserved."

Natural Language Understanding and Knowledge Graphs

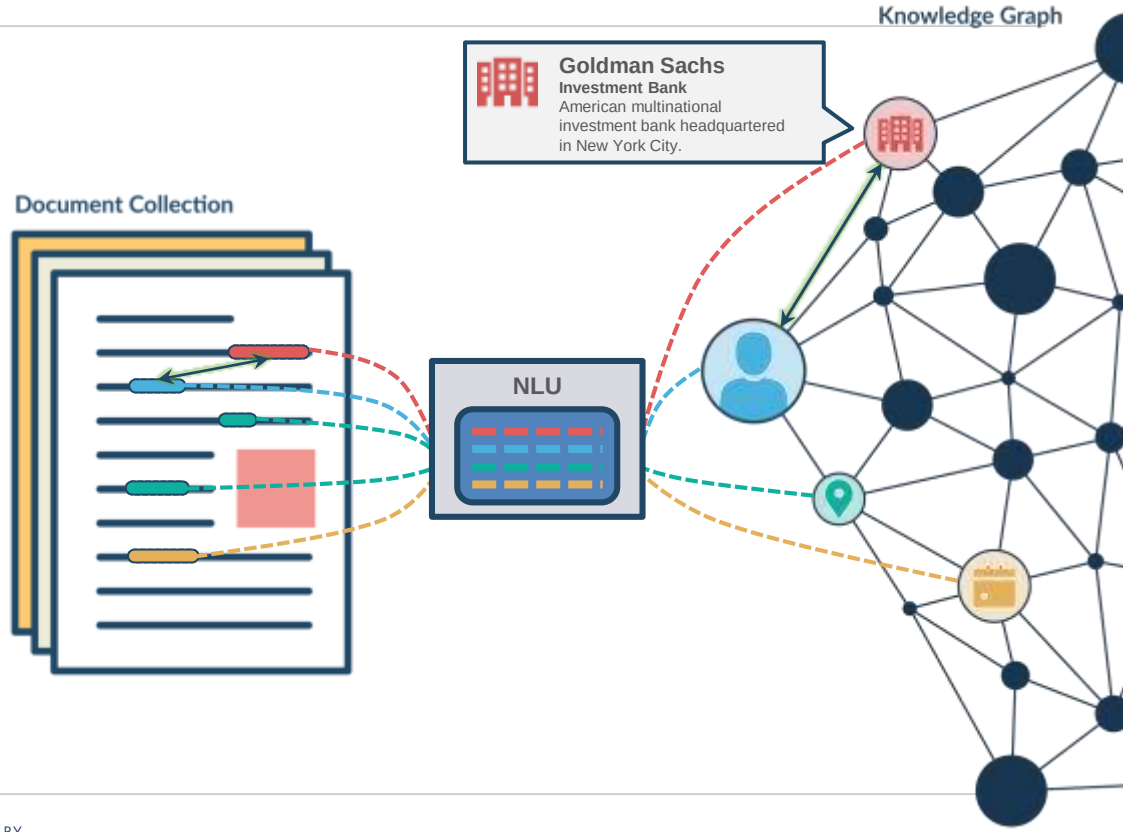
Journey of a Document

Discovers **entities** (people, organizations) and **concepts** (technical terms, jargon).

Links **entities** and **concepts** to knowledge graphs.

Identifies **relations** between the entities of interest.

Provides **background information** for entities and concepts.



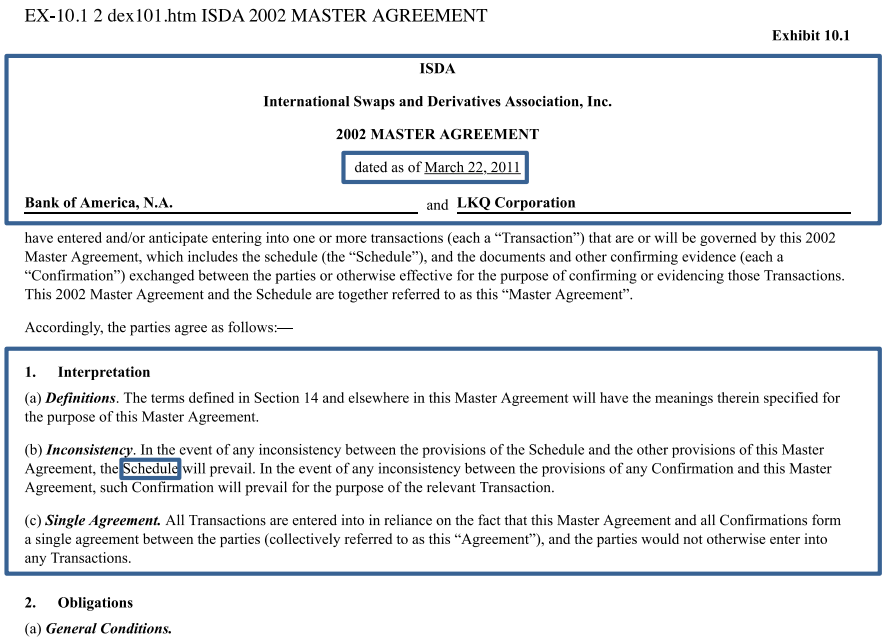
Icons made by Becirs, Madebyoliver, Freepik from www.flaticon.com is licensed by CC 3.0 BY

Finance Runs on Documents

NLP at GS

Finance Industry is built on data and information.

Financial Documents: Beyond Running Text



EX-10.1 2 dex101.htm ISDA 2002 MASTER AGREEMENT

Exhibit 10.1

ISDA

International Swaps and Derivatives Association, Inc.

2002 MASTER AGREEMENT

dated as of March 22, 2011

Bank of America, N.A. and **LKQ Corporation**

have entered and/or anticipate entering into one or more transactions (each a "Transaction") that are or will be governed by this 2002 Master Agreement, which includes the schedule (the "Schedule"), and the documents and other confirming evidence (each a "Confirmation") exchanged between the parties or otherwise effective for the purpose of confirming or evidencing those Transactions. This 2002 Master Agreement and the Schedule are together referred to as this "Master Agreement".

Accordingly, the parties agree as follows:—

1. Interpretation

(a) ***Definitions***. The terms defined in Section 14 and elsewhere in this Master Agreement will have the meanings therein specified for the purpose of this Master Agreement.

(b) ***Inconsistency***. In the event of any inconsistency between the provisions of the Schedule and the other provisions of this Master Agreement, the Schedule will prevail. In the event of any inconsistency between the provisions of any Confirmation and this Master Agreement, such Confirmation will prevail for the purpose of the relevant Transaction.

(c) ***Single Agreement.*** All Transactions are entered into in reliance on the fact that this Master Agreement and all Confirmations form a single agreement between the parties (collectively referred to as this "Agreement"), and the parties would not otherwise enter into any Transactions.

2. Obligations

(a) ***General Conditions.***

Also: credit agreements, prospectus, filings, …

1. Visually Structured
2. Deeply Hierarchical Content
3. Hyperlinks for Humans
4. Temporal Text

Beyond Running Text: Examples

Challenges

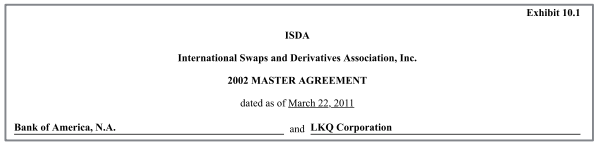


Exhibit 10.1

ISDA

International Swaps and Derivatives Association, Inc.

2002 MASTER AGREEMENT

dated as of March 22, 2011

Bank of America, N.A. and LKQ Corporation

Visually structured. Text will be arranged to facilitate human understanding. Formatted to preserve layout. Full of boxes and tables.

(n) Investments that do not exceed the sum of (x) the greater of (A) $100,000,000 and (B) 3.0% of Total Assets as of the end of the Test Period last ended at any time outstanding, plus (y) the Cumulative Amount at the time of such Investment;

Deeply hierarchical content. Section, subsection, paragraph, sub-paragraph, sentence hierarchy, …

(a) (i) Liens pursuant to any Loan Document and (ii) Liens on cash or deposits granted in favor of the Swing Line Lender or the L/C Issuer to Cash Collateralize any Defaulting Lender's participation in Letters of Credit or Swing Line Loans, respectively, as contemplated by Section 2.03(a)(ii)(E) and 2.04(b), and 2.17(a)(ii), respectively;

Hyperlinks for humans. Defined Terms and (internal and external) Sections explicitly referenced.

provided, further that for purposes of determining the Consolidated ~~Secured~~Total Net Debt Ratio for purposes of ~~Sections~~Section 2.14 ~~and 7.02(b)(20)~~ only, the cash proceeds of any Incremental Term Loan~~,~~ and/or Revolving Commitment Increase ~~and/or Permitted Debt Offering~~ shall not be deemed to be included on the consolidated balance sheet of the Borrower and its Restricted Subsidiaries.

Temporal Text. Inline strikethrough and edits, frequent amendments.

Everything is "English" language: constraints, logical expressions, statistical observations, …

Understanding structure of documents

Pages, Paragraphs, Lines, Words

THE GOLDMAN SACHS GROUP, INC. AND SUBSIDIARIES

Notes to Consolidated Financial Statements (Unaudited)

Loans and Lending Commitments

The table below presents the difference between the aggregate fair value and the aggregate contractual principal amount for loans and long-term receivables for which the fair value option was elected.

In the table above, the aggregate contractual principal amount of loans on nonaccrual status and/or more than 90 days past due (which excludes loans carried at zero fair value and considered uncollectible) exceeds the related fair value primarily because the firm regularly purchases loans, such as distressed loans, at values significantly below the contractual principal amounts.

The fair value of unfunded lending commitments for which the fair value option was elected was a liability of $38 million as of March 2019 and $45 million as of December 2018, and the related total contractual amount of these lending commitments was $7.27 billion as of March 2019 and $7.72 billion as of December 2018. See Note 18 for further information about lending commitments.

Long-Term Debt Instruments

The difference between the aggregate contractual principal amount and the related fair value of long-term other secured financings for which the fair value option was elected was not material as of both March 2019 and December 2018. The aggregate contractual principal amount of unsecured long-term borrowings for which the fair value option was elected exceeded the related fair value by $2.96 billion as of March 2019 and $3.47 billion as of December 2018. The amounts above include both principal-protected and non-principal-protected long-term borrowings.

Impact of Credit Spreads on Loans and Lending Commitments

The estimated net gain attributable to changes in instrument-specific credit spreads on loans and lending commitments for which the fair value option was elected was $77 million for the three months ended March 2019 and $108 million for the three months ended March 2018. The firm generally calculates the fair value of loans and lending commitments for which the fair value option is elected by discounting future cash flows at a rate which incorporates the instrument-specific credit spreads. For floating-rate loans and lending commitments, substantially all changes in fair value are attributable to changes in instrument-specific credit spreads, whereas for fixed-rate loans and lending commitments, changes in fair value are also attributable to changes in interest rates.

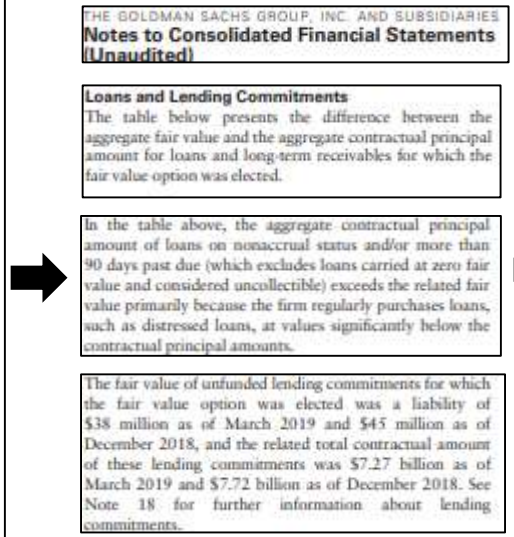
Debt Valuation Adjustment

The firm calculates the fair value of financial liabilities for which the fair value option is elected by discounting future cash flows at a rate which incorporates the firm's credit spreads.

The table below presents information about the net DVA gains/(losses) on financial liabilities for which the fair value option was elected.

In the table above:

- DVA (net of tax) is included in debt valuation adjustment in the consolidated statements of comprehensive income.
- The gains/(losses) reclassified to earnings from accumulated other comprehensive income/(loss) upon extinguishment of such financial liabilities were not material for both the three months ended March 2019 and March 2018.



THE GOLDMAN SACHS GROUP, INC. AND SUBSIDIARIES

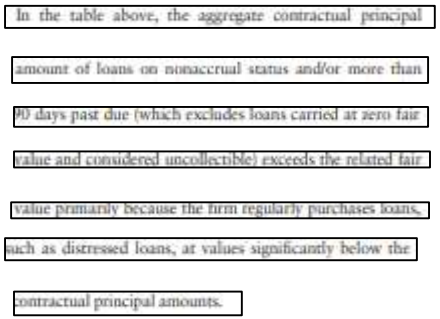
Notes to Consolidated Financial Statements (Unaudited)

Loans and Lending Commitments

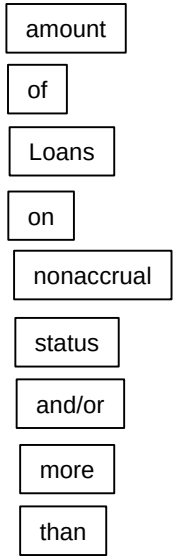
The table below presents the difference between the aggregate fair value and the aggregate contractual principal amount for loans and long-term receivables for which the fair value option was elected.

In the table above, the aggregate contractual principal amount of loans on nonaccrual status and/or more than 90 days past due (which excludes loans carried at zero fair value and considered uncollectible) exceeds the related fair value primarily because the firm regularly purchases loans, such as distressed loans, at values significantly below the contractual principal amounts.

The fair value of unfunded lending commitments for which the fair value option was elected was a liability of $38 million as of March 2019 and $45 million as of December 2018, and the related total contractual amount of these lending commitments was $7.27 billion as of March 2019 and $7.72 billion as of December 2018. See Note 18 for further information about lending commitments.



In the table above, the aggregate contractual principal amount of loans on nonaccrual status and/or more than 90 days past due (which excludes loans carried at zero fair value and considered uncollectible) exceeds the related fair value primarily because the firm regularly purchases loans, such as distressed loans, at values significantly below the contractual principal amounts.



amount of Loans on nonaccrual status and/or more than

Document representation for machines : Document Model

- Document Representation which can be understood by code
- Should be able to recover, represent and exploit visual structure within the document
- Should not lose any metadata information that is present in the document and potentially different for different document formats (.pdf, .html, .xlsx, etc.)
- Powerful enough to capture any aspect of the visual representation that is necessary for downstream knowledge extraction tasks

Using Document Model for Table Of Contents Extraction

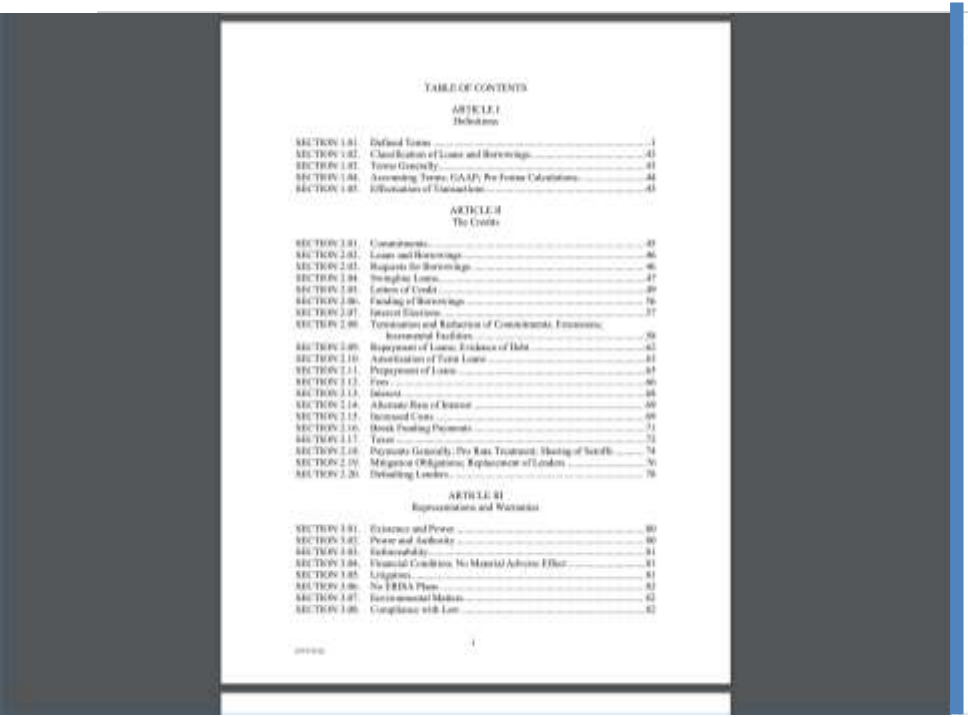


TABLE OF CONTENTS

ARTICLE I
Definitions

SECTION 1.01. Defined Terms 1
SECTION 1.02. Classification of Loans and Borrowings 43
SECTION 1.03. Terms Generally 43
SECTION 1.04. Accounting Terms; GAAP; Pro Forma Calculations 44
SECTION 1.05. Effectuation of Transactions 45

ARTICLE II
The Credits

SECTION 2.01. Commitments 45
SECTION 2.02. Loans and Borrowings 46
SECTION 2.03. Requests for Borrowings 46
SECTION 2.04. Swingline Loans 47
SECTION 2.05. Letters of Credit 49
SECTION 2.06. Funding of Borrowings 56
SECTION 2.07. Interest Elections 57
SECTION 2.08. Termination and Reduction of Commitments; Extensions of Incremental Facilities 58
SECTION 2.09. Repayment of Loans; Evidence of Debt 62
SECTION 2.10. Amortization of Term Loans 63
SECTION 2.11. Prepayment of Loans 65
SECTION 2.12. Fees 68
SECTION 2.13. Interest 69
SECTION 2.14. Alternate Rate of Interest 69
SECTION 2.15. Increased Costs 69
SECTION 2.16. Break Funding Payments 71
SECTION 2.17. Taxes 72
SECTION 2.18. Payments Generally; Pro Rata Treatment; Sharing of Setoffs 74
SECTION 2.19. Mitigation Obligations; Replacement of Lenders 76
SECTION 2.20. Defaulting Lenders 78

ARTICLE III
Representations and Warranties

SECTION 3.01. Existence and Power 80
SECTION 3.02. Power and Authority 80
SECTION 3.03. Enforceability 81
SECTION 3.04. Financial Condition; No Material Adverse Effect 81
SECTION 3.05. Litigation 81
SECTION 3.06. No ERISA Plans 82
SECTION 3.07. Environmental Matters 82
SECTION 3.08. Compliance with Law 82

Table Of Contents

- ARTICLE I Definitions
- SECTION 1.01 Defined Terms
- SECTION 1.02 Classification of Loans and Borrowings
- SECTION 1.03 Terms Generally
- SECTION 1.04 Accounting Terms; GAAP; Pro Forma Calculations
- SECTION 1.05 Effectuation of Transactions
- ARTICLE II The Credits
- SECTION 2.01 Commitments
- SECTION 2.02 Loans and Borrowings
- SECTION 2.03 Requests for Borrowings
- SECTION 2.04 Swingline Loans
- SECTION 2.05 Letters of Credit
- SECTION 2.06 Funding of Borrowings
- SECTION 2.07 Interest Elections
- SECTION 2.08 Termination and Reduction of Commitments; Extensions of Incremental Facilities
- SECTION 2.09 Repayment of Loans; Evidence of Debt
- SECTION 2.10 Amortization of Term Loans
- SECTION 2.11 Prepayment of Loans
- SECTION 2.12 Fees
- SECTION 2.13 Interest
- SECTION 2.14 Alternate Rate of Interest
- SECTION 2.15 Increased Costs
- SECTION 2.16 Break Funding Payments
- SECTION 2.17 Taxes
- SECTION 2.18 Payments Generally; Pro Rata Treatment; Sharing of Setoffs
- SECTION 2.19 Mitigation Obligations; Replacement of Lenders
- SECTION 2.20 Defaulting Lenders
- ARTICLE III Representations and Warranties
- SECTION 3.01 Existence and Power
- SECTION 3.02 Power and Authority

TABLE OF CONTENTS

ARTICLE I
Definitions

SECTION 1.01. Defined Terms 1
SECTION 1.02. Classification of Loans and Borrowings 43
SECTION 1.03. Terms Generally 43
SECTION 1.04. Accounting Terms; GAAP; Pro Forma Calculations 44
SECTION 1.05. Effectuation of Transactions 45

ARTICLE II
The Credits

SECTION 2.01. Commitments 45
SECTION 2.02. Loans and Borrowings 46
SECTION 2.03. Requests for Borrowings 46
SECTION 2.04. Swingline Loans 47
SECTION 2.05. Letters of Credit 49
SECTION 2.06. Funding of Borrowings 56
SECTION 2.07. Interest Elections 57
SECTION 2.08. Termination and Reduction of Commitments; Extensions of Incremental Facilities 58
SECTION 2.09. Repayment of Loans; Evidence of Debt 62
SECTION 2.10. Amortization of Term Loans 63
SECTION 2.11. Prepayment of Loans 65
SECTION 2.12. Fees 68
SECTION 2.13. Interest 69
SECTION 2.14. Alternate Rate of Interest 69
SECTION 2.15. Increased Costs 69
SECTION 2.16. Break Funding Payments 71
SECTION 2.17. Taxes 72
SECTION 2.18. Payments Generally; Pro Rata Treatment; Sharing of Setoffs 74
SECTION 2.19. Mitigation Obligations; Replacement of Lenders 76
SECTION 2.20. Defaulting Lenders 78

ARTICLE III
Representations and Warranties

SECTION 3.01. Existence and Power 80
SECTION 3.02. Power and Authority 80
SECTION 3.03. Enforceability 81
SECTION 3.04. Financial Condition; No Material Adverse Effect 81
SECTION 3.05. Litigation 81
SECTION 3.06. No ERISA Plans 82
SECTION 3.07. Environmental Matters 82
SECTION 3.08. Compliance with Law 82

Table of Contents

Hierarchy Extraction

Original Text

(t) other Liens securing obligations in an aggregate amount at any time outstanding not to exceed the greater of (x) $50,000,000, (y) 7.5% of Consolidated Total Assets (determined at the time of incurrence) and (z) the lesser of (1) 50% of Consolidated Adjusted EBITDA (for the most recent four fiscal quarter period for which financial statements have been delivered pursuant to Section 5.1(a) or Section 5.1(b) determined at the time of incurrence) and (2) $100,000,000.

Stanza Graph

(t) other Liens securing obligations in an aggregate amount at any time outstanding not to exceed the greater of
(x) $50,000,000,
(y) 7.5% of Consolidated Total Assets (determined at the time of incurrence) and
(z) the lesser of
(1) 50% of Consolidated Adjusted EBITDA (for the most recent four fiscal quarter period for which financial statements have been delivered pursuant to Section 5.1(a) or Section 5.1(b) determined at the time of incurrence) and
(2) $100,000,000.

use visual and semantic clues to derive a **hierarchical Stanza Structure**

Section and Sub Section References Extraction

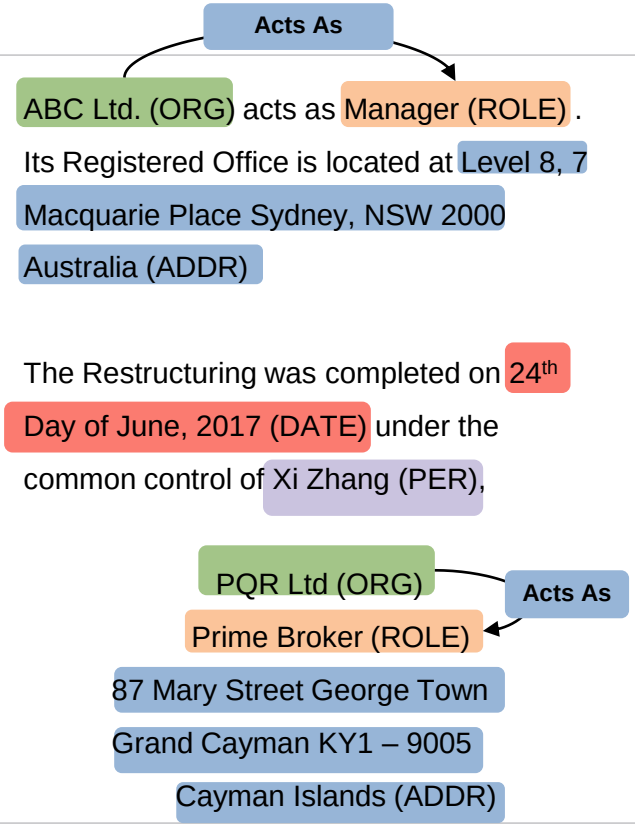
Hyperlinks to Section References and Defined Terms

From Structure to Knowledge

ABC Ltd. acts as the Manager. Its registered Office is located at Level 8, 7 Macquarie Place Sydney, NSW 2000 Australia

The Restructuring was completed on 24th Day of June, 2017 under the common control of Xi Zhang.

PQR Ltd
Prime Broker
87 Mary Street George Town
Grand Cayman KY1 – 9005
Cayman Islands



Extraction of Entities

Overview

The Fund invests its assets in ABC International Limited (\"ACIL\") a Delaware limited partnership. Mr. Jane acts as a director of XYZ Private Ltd.

Fine tuned Spacy
Post Processing Heuristics

The Fund invests its assets in ABC International Limited (ORG) (\"ACIL\") a Delaware limited partnership. Mr. Jane (PER) acts as a director of XYZ Private Ltd (ORG) .

The Registered Office of ABC Special Opportunities LTD. Is located at Level 8, 7 Macquarie Place Sydney, NSW 2000 Australia

NER with BERT backbone

The Registered Office of ABC Special Opportunities LTD. is located at Level 8, 7 Macquarie Place Sydney(City), NSW(State) 2000(Post Code) Australia(Country) (ADDR)

Extraction of Entities

Experimental Results

Organization Detection	Spacy large default model	Fined Tuned Spacy + Heuristics
F-score	0.368	**0.925**
Precision	0.32	**0.92**
Recall	0.43	**0.93**

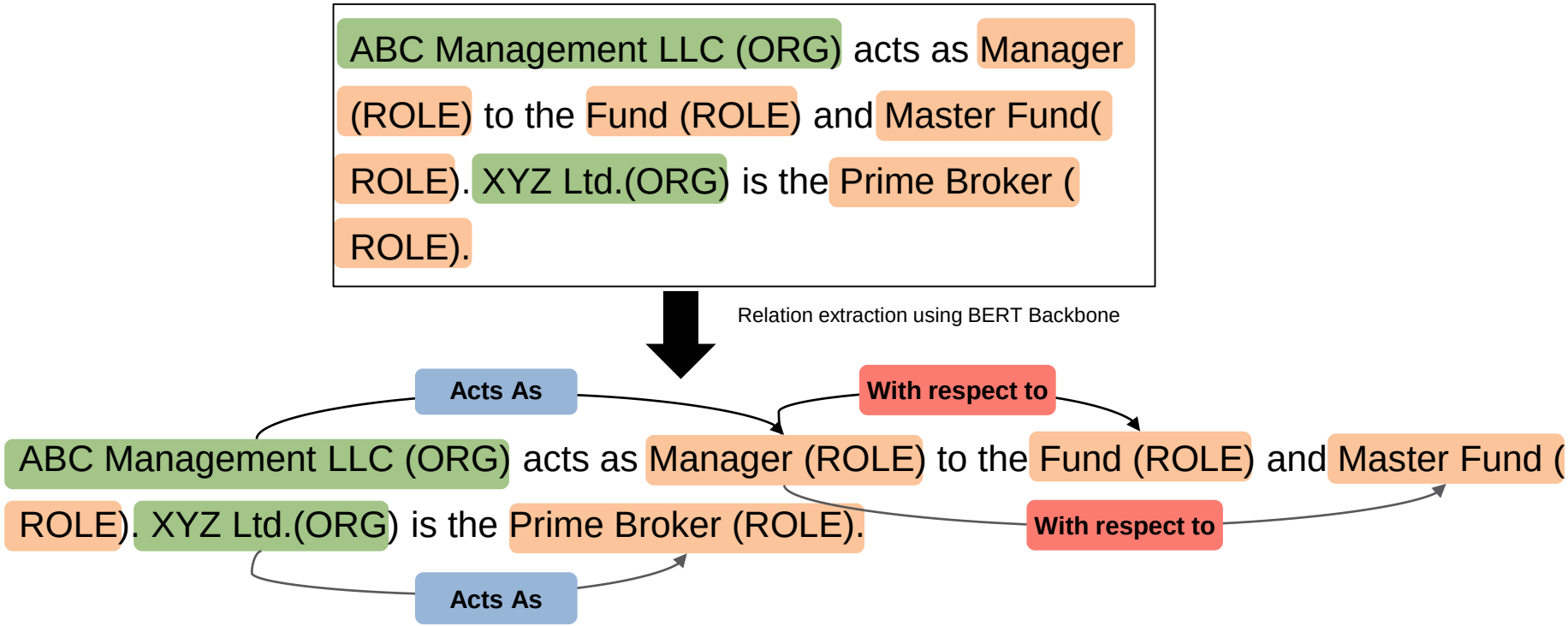
Person Detection	Spacy large default model	Fined Tuned Spacy
F-score	0.25	**0.94**
Precision	0.23	**0.91**
Recall	0.28	**0.96**

Address Span Detection	BERT	B Token	I Token	O Token
F-score	**0.959**	0.964	0.975	0.99
Precision	**0.941**	0.947	0.961	0.99
Recall	**0.978**	0.981	0.989	0.99

The BIO format (short for beginning, inside, outside) is a common format for tagging tokens in a chunking task like Named Entity Recognition

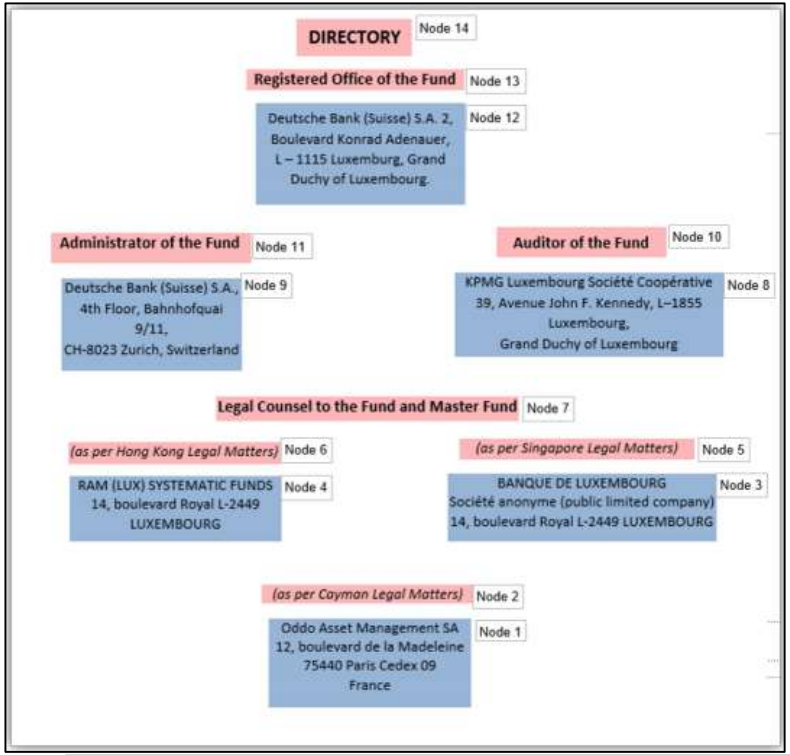
Relation Extraction

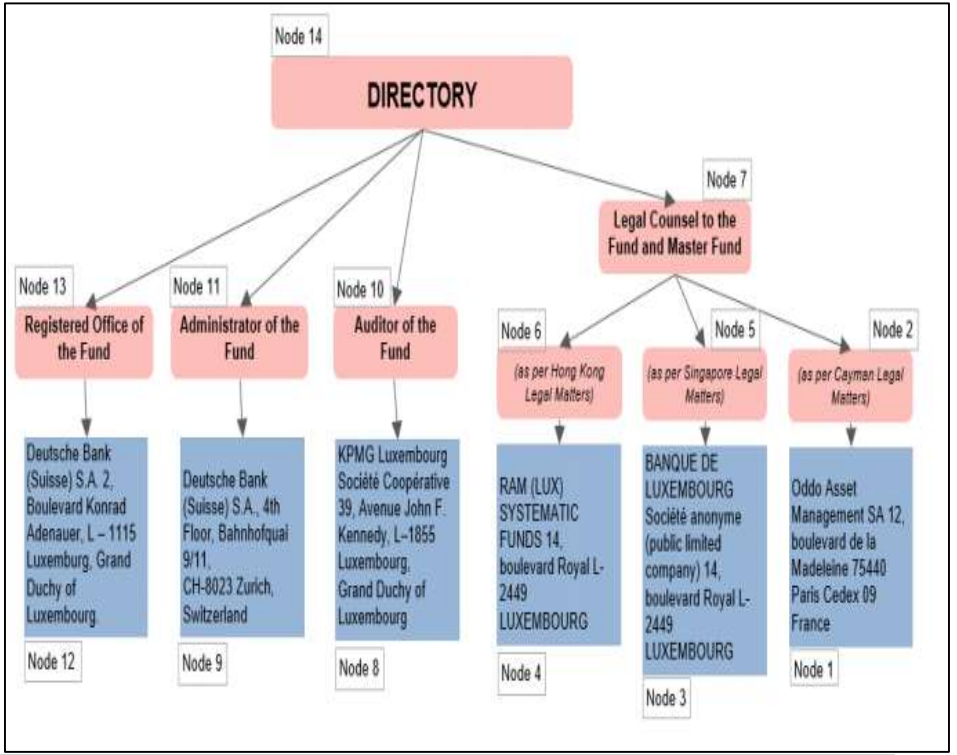
Overview



Combining Document Structure and Semantics

Parsing Directory Pages





Relation extraction and Directory Page Parsing

Experimental Results

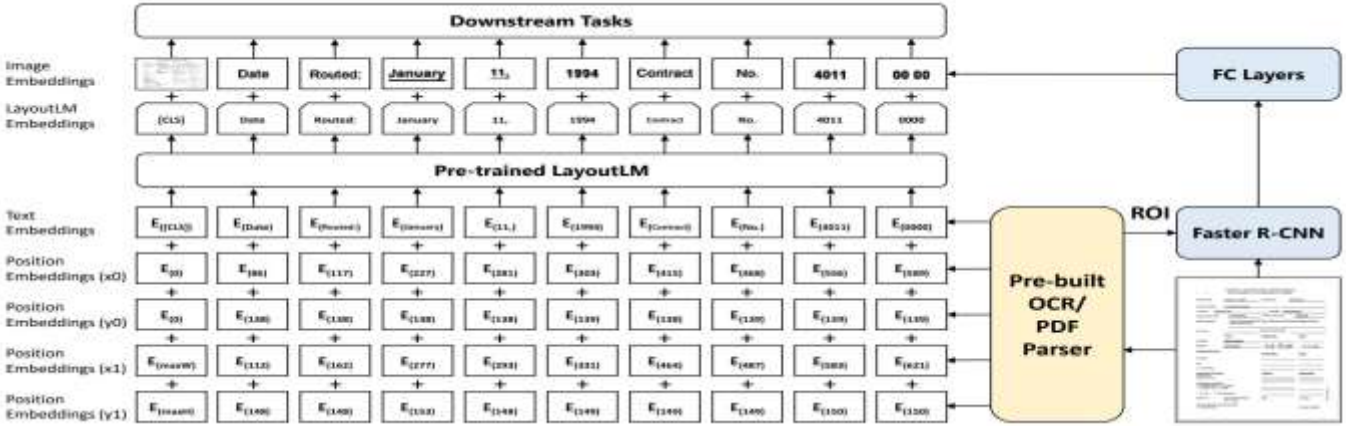
Alias Detection with respect to an entity	BERT + Entity markers
F-score	0.930
Precision	0.942
Recall	0.921

Recognizing Directory Pages	Random Forest
F-score	0.87
Precision	0.8
Recall	0.94

Determining Reading Order	Text Segmentation into Header and Body Nodes	Body Node and Its Immediate Parent	Dir Block
F-score	0.983	0.98	0.88
Precision	0.985	0.98	0.89
Recall	0.981	0.97	0.88

Future directions

Embeddings which respect both the semantics and the structure



Xu et al.: LayoutLM: Pre-training of Text and Layout for Document Image Understanding, KDD 2020
https://arxiv.org/abs/1912.13318

Concluding remarks

1. General open source tools are good for benchmarking but they are not fine tuned on your domain.

2. Annotate your own dataset – pays off well in the future

3. Optimise for a balance between inference time, memory consumed and accuracy

4. Distinguishing between the type of data – Textual versus Tabular.
   - Visual cues along with semantics have the potential to deliver powerful results

5. BERT is the beginning, but not the end. Transformer-based architectures keep on evolving and we need to stay on top of them to get the best performance

6. Build reusable tools

Goldman Sachs

QUESTIONS